MICHAŁ JAMRÓZ

# STRUKTURA I DYNAMIKA BIAŁEK GLOBULARNYCH – MODELOWANIE METODAMI GRUBOZIARNISTYMI


Praca doktorska wykonana w Pracowni Teorii Biopolimerów
Wydziału Chemii Uniwersytetu Warszawskiego

promotor: prof. dr hab. Andrzej Koliński

Warszawa, 2013

LABORATORY of THEORY of BIOPOLYMERS

PRACOWNIA TEORII BIOPOLIMERÓW
WYDZIAŁ CHEMII, UNIWERSYTET WARSZAWSKI
http://biocomp.chem.uw.edu.pl

# PODZIĘKOWANIA

Składam podziękowania wszystkim tym, którzy zechcieli przeczytać niniejszą pracę. Bez Czytelnika pisanie jej nie miałoby większego sensu.

Project operated within the Foundation for Polish Science MPD Programme MPD/2008/1 *"International Scholarship Program for Graduate Studies in Faculty of Chemistry University of Warsaw"* co-financed by the EU European Regional Development Fund. Economical support from the PhD grant number N N204 116539 is gratefully acknowledged.

# SPIS TREŚCI

# SPIS SKRÓTÓW

| | |
|---|---|
| CABS | Algorytm służący do modelowania struktur białek, wykorzystujący potencjały statystyczne i model gruboziarnisty białka |
| CASP | Międzynarodowy eksperyment sprawdzający postęp w rozwoju metod przewidywania struktur białek, ang.: *The Critical Assessment of Protein Structure Prediction* |
| CATH | Baza danych klasyfikująca struktury białek pod względem ich klasy, architektury, topologii i rodzin homologicznych. ang.: *Class, Architecture, Topology, Homologous superfamily* |
| CPU | Procesor komputera, ang.: *Central Processing Unit* |
| ENM | Model sieci elastycznej, ang.: *Elastic Network Model* |
| FRET | Metoda, w której mierzone jest widmo emisyjne przeniesienia energii pomiędzy dwoma chromoforami, ang.: *Förster Resonance Energy Transfer* |
| GNM | Model sieci Gaussowskiej, ang.: *Gaussian Network Model* |
| GPU | Procesor karty graficznej, ang.: *Graphics Processing Unit* |
| HTTP | Protokół przesyłania dokumentów hipertekstowych w sieci internet, ang.: *HyperText Transfer Protocol* |
| MD | Dynamika molekularna, ang.: *Molecular Dynamics* |
| MODELLER | Algorytm służący do modelowania białek, wykorzystujący pole siłowe wykorzystujące m.in. więzy odległości wzięte ze struktur białek-szablonów |
| NMA | Analiza drgań normalnych, ang.: *Normal Mode Analysis* |

| | |
|---|---|
| NMR | Magnetyczny rezonans jądrowy, ang.: *Nuclear Magnetic Resonance* |
| NOE | Efekt jądrowy Overhausera, ang.: *Nuclear Overhauser Effect* |
| PDB | Baza danych doświadczalnie rozwiązanych struktur białek, ang.: *Protein Data Bank* (Berman i in., 2000) |
| RDC | Resztkowe sprzężenia dipolowe, ang.: *Residual Dipolar Couplings* |
| REMC | Metoda wymiany replik Monte Carlo, ang.: *Replica Exchange Monte Carlo* |
| REMD | Metoda wymiany replik w dynamice molekularnej, ang.: *Replica Exchange Molecular Dynamics* |
| RMSD | Pierwiastek średniego kwadratowego odchylenia par atomów po optymalnym nałożeniu, ang.: *Root Mean Square Deviation*, szczegółowo opisane w Dodatku G.3 |
| ROSETTA | Algorytm służący do modelowania białek, wykorzystujący bazę danych fragmentów struktur białkowych |
| SAXS | Małokątowe rozpraszanie promieni rentgenowskich, ang.: *Small Angle X-ray Scattering* |
| SVR | Regresja z wykorzystaniem maszyn wektorów nośnych, ang.: *Support Vector Regression*, Dodatek F |
| TROSY | Spektroskopia optymalizowana pod kątem pomiarów relaksacji poprzecznej. ang.: *Transverse Relaxation Optimised SpectroscopY* |

# Część I

# Wprowadzenie

# 1 WSTĘP

Funkcjonowanie organizmów żywych opiera się na mechanizmach, których ważną częścią są białka, wielkocząsteczkowe związki o złożonej strukturze. Białka uczestniczą praktycznie we wszystkich procesach zachodzących w organizmie, między innymi przy przekazywaniu sygnałów między komórkami, transporcie związków wewnątrz komórek i katalizowaniu reakcji metabolicznych.

Wiele obecnie stosowanych leków opiera się na blokowaniu, bądź stymulacji określonego białka czy grupy białek. Jednym z przykładów takiego działania może być kwas acetylosalicylowy (*aspiryna*), działający przeciwgorączkowo, przeciwbólowo i przeciwzapalnie. Mechanizm jego działania polega na inhibicji cyklooksygenazy, enzymu katalizującego proces powstawania prostaglandyn, które wywołują m.in. reakcję zapalną w organizmie (Vane, 1971). Znając strukturę cyklooksygenazy możliwe jest nie tylko wyjaśnienie reakcji inhibicji enzymu na poziomie atomowym (Kaur i in., 2012), ale również projektowanie nowych leków o podobnym do kwasu acetylosalicylowego działaniu (Mozziconacci i in., 2005).

## 1.1 FUNKCJA BIAŁKA A JEGO STRUKTURA

Anfinsen (1973) pokazał, że sekwencja aminokwasowa białka determinuje jego strukturę trzeciorzędową (strukturę posiadająca minimalną wartość energii swobodnej Gibbsa), powstającą w procesie zwijania (fałdowania). Struktura ta natomiast determinuje funkcję poprzez dynamikę zwiniętego łańcucha.

Obecnie schemat sekwencja $\rightarrow$ struktura $\rightarrow$ funkcja uzupełnia się członem „dynamika" tj. sekwencja $\rightarrow$ struktura $\rightarrow$ dynamika $\rightarrow$ funkcja (Bahar i Rader, 2005), bowiem białka są nie tylko obiektami statycznymi, ale również dyna-

micznymi (Huang i Montelione, 2005). Aby w pełni zrozumieć ich funkcję należy badać je na przykład w domenie czasu (Eisenmesser i in., 2002) lub na podstawie zbioru konformacji (Kmiecik i Kolinski, 2007).

Rasmussen i in. (1992) pokazali, że drgania atomów są niezbędne dla prawidłowej pracy enzymu rybonukleazy A: enzym ten traci funkcję, wykazując harmoniczne drgania poszczególnych atomów, w temperaturze poniżej 220 K; zaś powyżej tej temperatury drgania termiczne zdominowane są przez anharmoniczne ruchy kolektywne (wspólne ruchy większych fragmentów białka). Z kolei Eisenmesser i in. (2005) pokazali, że kolektywne fluktuacje cyklofiliny A (CypA) występujące podczas reakcji enzymatycznej obecne są również w strukturze bez substratu, co sugeruje, że taka specyficzna dynamika łańcucha jest prawdopodobnie naturalną cechą cyklofiliny A, a zarazem jest niezbędna do jej poprawnego działania.

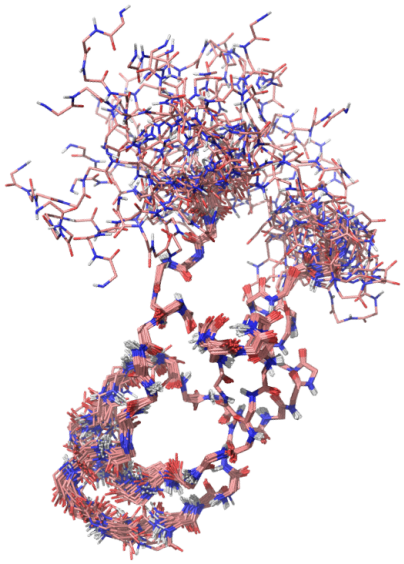
Energia swobodna Gibbsa, $G = H - TS$, wyraża energię całkowitą układu. Entalpię i entropię można nazwać, kolejno, komponentami statycznymi i dynamicznymi energii układu (Homans, 2005). Komponent statyczny wyraża się przez modelowe potencjały danego układu, komponent dynamiczny nastręcza większych trudności, gdyż wyrażać musi również entropię otoczenia (rozpuszczalnika). Wydaje się, że lepsze zbadanie dynamiki (chociażby entropii wyrażonej przez zbiór konformacji) pozwoli na skonstruowanie skuteczniejszych modeli wiązania liganda z receptorem, co z kolei pozwoli na efektywniejsze projektowanie leków *in silico*[1].

Wykorzystując informację o dynamice białek bada się m.in. mechanizmy ich funkcjonowania (Teilum i in., 2009), reakcje enzymatyczne (Hammes i in., 2011), proteopatie (Chiti i Dobson, 2009) czy mechanizmy wzajemnego oddziaływania pomiędzy nimi (Zacharias, 2010). Wiedzę o dynamice białka stosuje się również przy projektowaniu nowych enzymów (Mandell i Kortemme, 2009; Lassila, 2010) oraz leków (Lill, 2011).

Historia ewolucji modelu łączenia się liganda z receptorem może być przykładem tego, jak z biegiem lat coraz większą uwagę przywiązywano do aspektu dynamiki białek: jednym z pierwszych proponowanych modeli oddziaływania enzymu z substratem był mechanizm „klucza i zamka“, zaproponowany pod koniec XIX wieku przez Fishera (Fischer, 1894). W modelu tym

1 Z użyciem komputera.

enzym posiadał miejsce dobrze dopasowane strukturalnie i niejako „czekające" na specyficzny substrat. W miarę postępu badań wiele enzymów nie dawało się opisać w ten sposób, dlatego też model ten został uzupełniony o model indukowanego dopasowania, gdzie centrum wiążące enzymu zmienia konformację w miarę zbliżania się substratu, tak by lepiej się z nim związać (Koshland, 1958; Ma i in., 1999). Obecnie zaś proponuje się tzw. selekcję konformacyjną (model Monoda-Wymana-Changeux opracowany na potrzeby opisu mechanizmów przejść allosterycznych (Monod i in., 1965)), gdzie enzym postrzegany jest nie jako pojedyncza struktura, a jako zbiór konformerów o podobnej energii swobodnej Gibbsa (Kern i Zuiderweg, 2003). Model ten później zastosowano z powodzeniem również do opisu enzymów „nieallosterycznych" (Gunasekaran i in., 2004). Tego typu opisy oddziaływań uwzględniają czynnik dynamiczny (entropię wyrażoną przez entropię konformacyjną).



**Rysunek 1:** Rozwiązana z użyciem spektroskopii NMR struktura jednej z domen białka TFPI (kod PDB: 1adz).

Innym — może bardziej obrazowym — przykładem konieczności uwzględniania ruchliwości łańcucha białkowego w trakcie badań może być białko inhibitora tromboplastyny tkankowej (TFPI) (Rysunek 1), które w stanie natywnym występuje raczej jako zbiór konformerów niż pojedyncza struktura. Przypuszczalnie, opieranie się na jednej tylko konformacji mogłoby prowadzić do błędów w interpretacji wyników (np. ocena poprawności modelu

przewidzianego teoretycznie z użyciem RMSD i przy zastosowaniu tylko jednej ze struktur białka z Rysunku 1 jako struktury odniesienia).

Choć przez większość wstępu odnosiłem się do dynamiki białek, należy pamiętać, że dynamika opisuje ruch struktury, w związku z czym badanie dynamiki implikuje badanie struktury.

## 1.2 CEL PRACY

Celem pracy opisanej w niniejszej rozprawie było zastosowanie gruboziarnistych metod modelowania komputerowego do przewidywania struktury i dynamiki białek.

W pracy porównałem wyniki otrzymane z metod zredukowanych z danymi otrzymanymi z użyciem klasycznej, pełno-atomowej dynamiki molekularnej. Wyniki pracy pokazały, że dla krótkich czasów symulacji (10 ns) gruboziarnisty model CABS jest konsystentny w opisie z dynamiką molekularną łańcucha białkowego, co pozwala na przeprowadzanie symulacji dynamiki stanu natywnego w znacznie krótszym czasie użycia procesora (w porównaniu z MD).

W pracy zbadałem wpływ różnych parametrów opisu struktury białka na ruchliwość atomów łańcucha głównego (atomów węgla C$\alpha$). Używając najlepszych zestawów parametrów stworzyłem model *maszyny wektorów nośnych* służący do przewidywania wartości fluktuacji atomów C$\alpha$ na podstawie samej struktury.

Zastosowałem podejście wieloskalowe do modelowania *de novo* fragmentów pętli w białkach, proponując uniwersalną metodykę modelowania pętli o długości do 25. reszt aminokwasowych. Rozszerzając to zagadnienie, opracowałem półautomatyczną metodę przewidywania struktury białek.

# Część II

# Cel rozprawy i opis wykorzystanych metod

# 2 STRUKTURA BIAŁEK

## 2.1 METODY DOŚWIADCZALNE WYZNACZANIA STRUKTUR BIAŁEK

Obecnie (RCSB Protein DataBank, 2012) znamy przeszło osiemdziesiąt tysięcy struktur białek rozwiązanych metodami doświadczalnymi — około 89% z użyciem rentgenografii strukturalnej oraz około 11% z wykorzystaniem spektroskopii NMR, co z jednej strony jest liczbą dość pokaźną uwzględniwszy wagę posiadania struktury wyznaczonej eksperymentalnie, z drugiej zaś strony jest to wciąż niewiele, gdy weźmie się pod uwagę liczbę sekwencji białkowych w bazie UniProtKB/TrEMBL — znajduje się tam ponad 30 milionów pozycji[1].

Należy jednak mieć na uwadze, że obecne metody doświadczalnego rozwiązywania struktur wciąż nie radzą sobie z przypadkami, gdzie niemożliwe jest otrzymanie monokryształu (większość białek membranowych lub białka w znacznym stopniu nieustrukturyzowane), oraz gdy specyfika białka nie pozwala na otrzymanie danych NMR możliwych do przetłumaczenia na współrzędne położeń atomów w strukturze.

Pomimo ciągle rosnącej liczby zdeponowanych struktur białek w bazie PDB, wiele z nich jest podobnych sekwencyjnie; natomiast struktura białek, których nie udaje się rozwiązać metodami rezonansu magnetycznego lub krystalograficznie wciąż pozostaje nieznana. Między innymi dlatego stosuje się komputerowe metody przewidywania struktury białek.

1 http://www.ebi.ac.uk/uniprot/TrEMBLstats/

## 2.2 METODY TEORETYCZNE WYZNACZANIA STRUKTUR BIAŁEK

Moc obliczeniowa współczesnych komputerów oraz stosowane pola siłowe pozwalają otrzymać struktury białek, których nie udało się rozwiązać doświadczalnie — zwykle w krótszym czasie i przy niższym nakładzie finansowym. Do tej pory powstało kilkadziesiąt algorytmów rozwiązywania struktur białek, z których bardziej popularnymi są:

- MODELLER (Sali i Blundell, 1993; Eswar i in., 2007): algorytm, który na podstawie m.in. więzów odległości branych ze znanych struktur podobnych sekwencyjnie do sekwencji celu tworzy pełno-atomowe modele, które następnie są optymalizowane odpowiednim potencjałem (np. DOPE (Shen i Sali, 2006)),

- ROSETTA (Rohl i in., 2004): algorytm, który tworzy zbiory modeli z użyciem fragmentów łańcuchów białkowych podobnych sekwencyjnie do sekwencji celu, po czym wybiera modele reprezentatywne w etapie oceny polem siłowym oraz z użyciem analizy skupień,

- modele gruboziarniste TASSER (Wu i in., 2007), UNRES (Liwo i in., 2005), CABS (Kolinski, 2004a; Kolinski i Skolnick, 2004), TOUCHSTONE (Skolnick i in., 2003; Kihara i in., 2002; Li i in., 2003; Zhang i in., 2003), w których pewne grupy atomów reprezentowane są przez *pseudo-atomy* (centra oddziaływań, często o wyłączonej objętości), pozwalając tym samym na znaczne zredukowanie kosztów obliczeń (próbkowanie większego obszaru przestrzeni konformacyjnej w tym samym czasie pracy procesora),

- dynamika molekularna (MD) z polami siłowym zoptymalizowanymi dla struktur białek, np. OPLS/AA (Jorgensen i Tirado-Rives, 1988) czy CHARMM (Brooks i in., 2009). Ze względu na kosztowność obliczeń nie stosuje się jej w praktyce do przewidywania struktury białka, choć pokazano kilka przykładów symulacji, gdzie startując z rozwiniętej konformacji krótkich białek otrzymuje się strukturę natywną (Lindorff-Larsen i in., 2011).

Większość z metod przewidywania struktury opiera się na założeniu, że ewolucyjnie zbliżone białka posiadają podobną sekwencję, a tym samym strukturę (Kaczanowski i Zielenkiewicz, 2009). Używając algorytmów dopasowania sekwencji, takich jak BLAST (Altschul i in., 1997), znajduje się w bazie rozwiązanych struktur modele-szablony, które służą za struktury początkowe w procesie modelowania — optymalizacji geometrii całości oraz tworzeniu *de novo* fragmentów, dla których w dopasowaniu sekwencyjnym pojawiały się przerwy.

W części praktycznej opiszę zastosowanie połączenia gruboziarnistego modelu CABS wraz z algorytmem MODELLER do modelowania *de novo* konformacji fragmentów pętli w strukturach białek. Jest to o tyle istotne, że przy badaniu dynamiki stanu natywnego czy w eksperymencie dokowania ligandów wymagane jest posiadanie kompletnego łańcucha białkowego. Zdarza się również, że same pętle pełnią czasami rolę centrów aktywnych (Lee i in., 2010). Modelowanie pętli jest również krytycznym testem pól siłowych, bowiem tworzenie kolejnych konformacji następuje z wykorzystaniem jedynie definicji pola siłowego oraz informacji o sekwencji aminokwasowej pętli i jej otoczenia. Wreszcie, generowane konformacje pętli mogą być stosowane jako przybliżona trajektoria ich dynamiki.

### 2.2.1 Eksperyment CASP jako ocena teoretycznych metod przewidywania struktury białek

Eksperyment CASP jest odbywającym się co dwa lata konkursem oceniającym rozwijane metody przewidywania struktury białek. Polega on na udostępnianiu uczestniczącym w nim grupom sekwencji białek bez udostępnienia struktur. Zespoły badawcze — używając własnych metod — starają się przewidzieć ich strukturę, dzięki czemu mogą porównać opracowywane przez siebie metody z metodami innych uczestników pod względem poprawności przewidzianej struktury.

Po każdym eksperymencie publikowane są raporty z postępów w dziedzinie przewidywania struktur białek. Porównując wyniki dziewiątej tury CASP z poprzednimi turami (Kryshtafovych i in., 2011) można zauważyć, że rozwój metod predykcji struktury osiągnął pewien kres i od około dziesięciu lat

poprawność predykcji metodami *in silico* utrzymuje się na mniej więcej stałym, dość dobrym poziomie; natomiast dla sekwencji modelowanych *de novo* (bez struktury-szablonu) można otrzymać rozsądne struktury dla łańcuchów o maksymalnej długości około 120. reszt aminokwasowych (Kryshtafovych i in., 2011).

Z eksperymentu CASP można wyciągnąć wniosek, że obecne metody predykcji struktury pozwalają na otrzymanie poprawnych struktur (lub przynajmniej fragmentów) dla większości sekwencji.

Warto nadmienić, że grupa *Kolinski-Bujnicki*, stosująca algorytm CABS w konkursie CASP6 została sklasyfikowana jako druga (lub pierwsza, zależnie od sposobu oceny dokładności modeli) spośród około dwustu najlepszych grup badawczych z całego świata (Debe i in., 2006; Koliński i Bujnicki, 2005).

## 2.3 OPIS STRUKTURY BIAŁKA

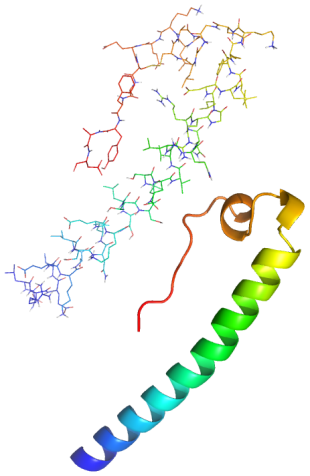
Strukturę białka można scharakteryzować zestawem parametrów opisujących ją na poziomie rozmiarów reszty aminokwasowej. Dzięki takiemu podejściu zmniejsza się złożoność opisu, zachowując (lub nieznacznie redukując) informację o strukturze pierwotnej.

I tak na przykład zamiast reprezentacji pełno-atomowej można zdefiniować pewne charakterystyczne wzorce wiązań wodorowych pomiędzy atomami łańcucha głównego, przedstawiając je jako $\alpha$-helisy czy $\beta$-wstęgi (tzw. struktura drugorzędowa), ułatwiając w ten sposób interpretację, co przedstawiłem na Rysunku 2.

W dalszej części rozdziału opiszę parametry statyczne (tj. statycznie opisujące białko), które wykorzystałem do zdefiniowania struktury i zbadania ich wpływu na mobilność łańcucha (Praca D).

### 2.3.1 Sekwencja aminokwasów

Sekwencja aminokwasów (tzw. struktura pierwszorzędowa) jako całość w sposób naturalny determinuje mobilność łańcucha białkowego. Opracowano kilka metod predykcji fluktuacji łańcucha na podstawie samej sekwencji




**Rysunek 2:** Reprezentacja pełno-atomowa (z pominięciem atomów wodorów) oraz reprezentacja przedstawiająca strukturę drugorzędową (zdefiniowaną jako pewne powtarzalne wzory wiązań wodorowych) białka kapsydu wirusa MoMuLV (kod PDB: 1mof)

(Schlessinger i Rost, 2005; Hirose i in., 2010; Bornot i in., 2011; Gu i in., 2006; Schlessinger i in., 2006; Pan i Shen, 2009), otrzymując narzędzia pozwalające na rozróżnienie fragmentów białka o większej/mniejszej mobilności.

Należy jednak zaznaczyć, że średnia wartość amplitud fluktuacji[2] poszczególnych aminokwasów jest podobna, co oznacza, że ruchliwość danego aminokwasu nie jest determinowana przez jego typ, a raczej otoczenie, w którym się znajduje (Smith i in., 2003). Chcąc zatem używać jedynie sekwencji aminokwasowej przy szacowaniu różnic w mobilności poszczególnych reszt, konieczne jest traktowanie jej (sekwencji) jako całości.

### 2.3.2 Struktura drugorzędowa

Najczęściej wykorzystywanym sposobem opisu struktury drugorzędowej jest ten zaproponowany w pracy Kabsch i Sander (1983). Sposób ten wykorzystuje osiem definicji wzorców wiązań wodorowych pomiędzy atomami łańcucha głównego: $\alpha$-helisy („H"), $\beta$-wstęgi („E"), $3_{10}$-helisy („G"), $\pi$-helisy („I"), $\beta$-mostka („B"), zwrotu („T"), zagięcia (ang.: *bend*, „S") oraz innego/pętli (bez oznaczenia, bądź „C"). Ze względu na częstość występowania, podział

2 Wyrażona poprzez czynnik temperaturowy struktur rozwiązanych z użyciem rentgenografii strukturalnej; szerzej opisany w Rozdziale 3.1.1, strona 20.

ten zwykle upraszcza się do reszt aminokwasowych typu $\alpha$-helisy („H") (włączając tu również reszty opisane jako $3_{10}$-helisy oraz $\pi$-helisy), $\beta$-wstęgi („E") oraz pozostałych („C").

Regiony nieustrukturyzowane wiązaniami wodorowymi (C) wykazują największą mobilność, elementy $\alpha$-helis pośrednią, natomiast najbardziej sztywne są elementy $\beta$-wstęg (średnie wartości znormalizowanego czynnika temperaturowego, odpowiednio, 0,27, -0,14, -0,37 Yuan i in. (2003), podobne wyniki przedstawili Zhang i in. (2009).

### 2.3.3 Powierzchnia wyeksponowana do rozpuszczalnika

Próbkując białko sferą o promieniu van der Waalsa rozpuszczalnika (wody) można określić powierzchnię poszczególnych reszt mających kontakt z rozpuszczalnikiem (Sanner i in., 1996). Tak zdefiniowany parametr dobrze rozróżnia reszty mobilne (znajdujące się na powierzchni białka) oraz reszty sztywne (składające się na rdzeń białka). Średni znormalizowany czynnik temperaturowy dla reszt zagrzebanych wynosi -0,508, natomiast dla reszt wyeksponowanych do rozpuszczalnika: 0,248 (Yuan i in., 2003). Wynika to głównie z gęstości upakowania atomów wewnątrz białka, która jest większa niż na powierzchni.

### 2.3.4 Odległość atomu od środka masy białka

Shih i in. (2007) zauważyli, że kwadrat odległości atomu od środka masy białka (Równanie 2.1) dobrze koreluje z czynnikami temperaturowymi modeli białek, aczkolwiek by otrzymać podobne wartości należy uzyskane $r_i^2$ przeskalować do zakresu czynników temperaturowych (innymi słowy tą metodą otrzymuje się względne wartości fluktuacji).

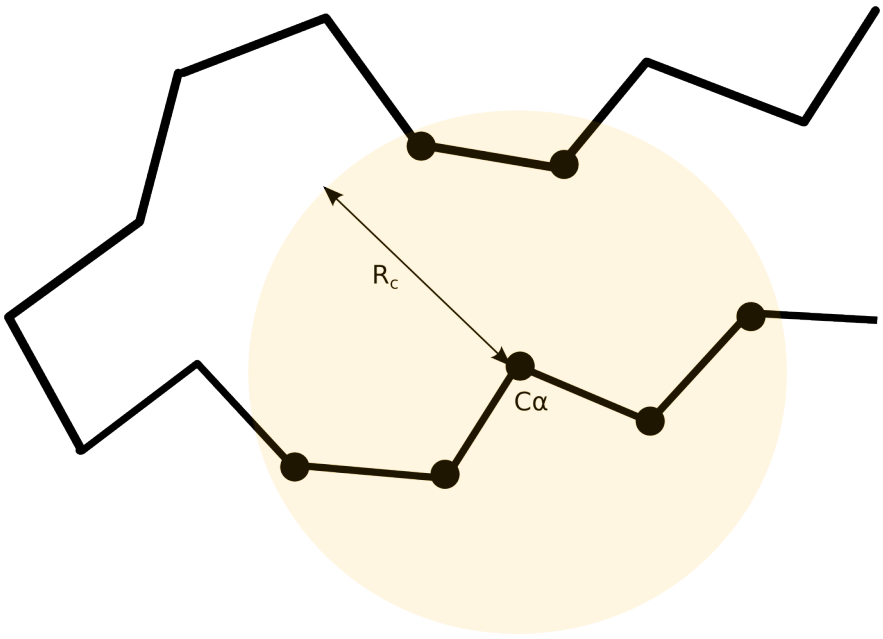
$$r_i^2 = \left( x_i - \frac{1}{N} \sum_j^N m_j x_j \right)^2 \tag{2.1}$$

gdzie:

| | |
|---|---|
| $N$ | liczba atomów w strukturze, |
| $m_j$ | masa atomu j, |
| $x_j$ | wektor położenia atomu j. |

### 2.3.5 Liczba kontaktów (liczba koordynacyjna)

Ostatnio (Halle, 2002; Lin i in., 2008) pokazano, że liczba kontaktów wokół atomu C$\alpha$ jest odwrotnie proporcjonalna do amplitudy fluktuacji.




**Rysunek 3:** Czarnymi punktami zaznaczyłem atomy C$\alpha$ wchodzące w kontakt z atomem centralnym (będące w promieniu odcięcia $R_c$). W tym przypadku $C_i = 6$.

Liczbę kontaktów definiuje się jako ilość otaczających dany atom innych atomów, które są bliżej niż zadany promień odcięcia $R_c$ (Rysunek 3):

$$C_i = \sum_{j \neq i}^{N} c_j \tag{2.2}$$

w którym:

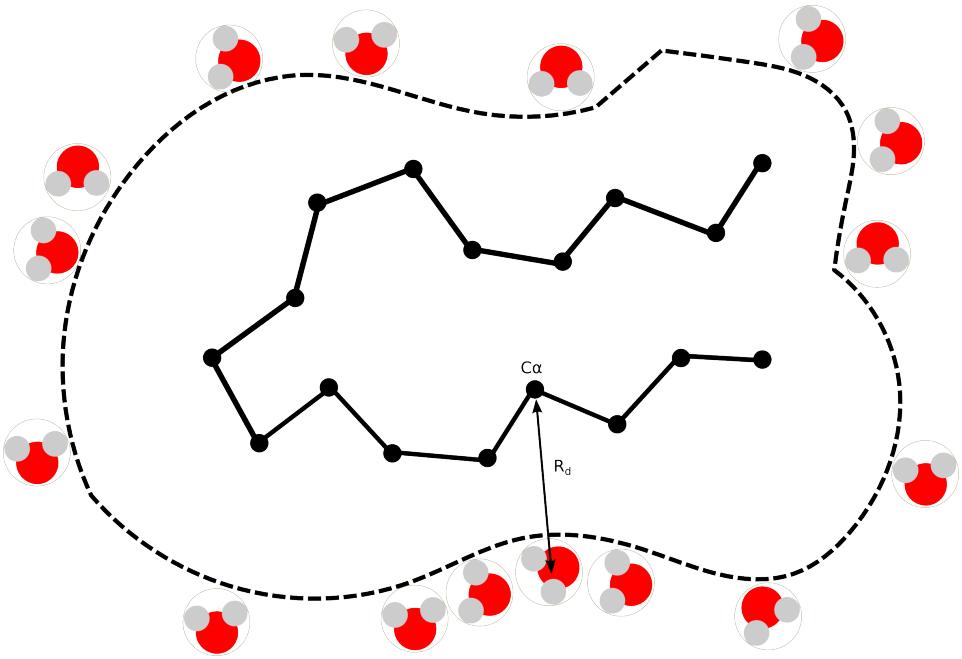
$$c_j = \begin{cases} 1 & \text{gdy} \quad d_{ij} \leqslant R_c \\ 0 & \text{gdy} \quad d_{ij} > R_c \end{cases} \tag{2.3}$$

gdzie:

$d_{ij}$ odległość między atomem i-tym, a j-tym.

### 2.3.6 Zanurzenie reszty aminokwasowej (ang.: *Residue depth*)

Parametr ten definiuje się jako odległość danego atomu (bądź środka masy grupy bocznej) do najbliższej cząsteczki wody przy powierzchni białka (Rysunek 4) (Chakravarty i Varadarajan, 1999).




**Rysunek 4:** Zanurzenie reszty w białku jako odległość ($R_d$) między atomem C$\alpha$, a najbliższą cząsteczką wody.

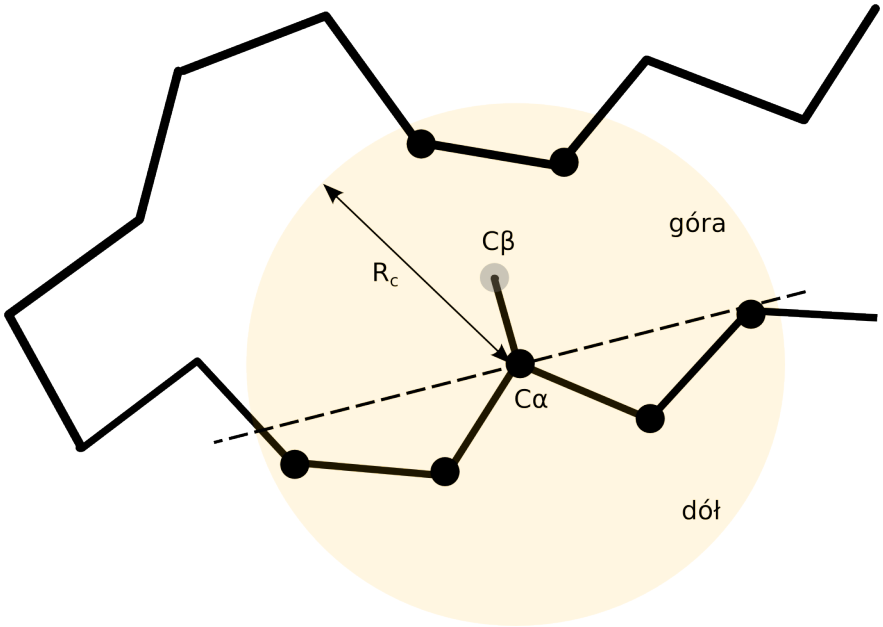
### 2.3.7 Otoczenie hydrofobowe/hydrofilowe reszty

Używając skali hydrofobowości Kyte i Doolittle (1982) i opisując otaczające dany atom reszty jako hydrofilowe lub hydrofobowe (ujemne lub dodatnie

wartości na skali Kyte-Doolittle) można przedstawić otoczenie atomu dwoma parametrami — liczbą kontaktów hydrofobowych i hydrofilowych.

### 2.3.8 Liczba kontaktów dolnej/górnej części sfery reszty aminokwasowej

Parametr zdefiniowany jako liczba kontaktów (Równanie 2.2) w obrębie dolnej/górnej części sfery o promieniu 13 Å od atomu Cα (lub Cβ), gdzie płaszczyzna podziału wyznaczona jest jako prostopadła do wektora łączącego atomy Cα i Cβ (Rysunek 5).




**Rysunek 5:** Czarnymi punktami zaznaczono atomy Cα wchodzące w kontakt (będące w promieniu odcięcia $R_c$) z atomem centralnym. Przerywaną linią oznaczono płaszczyznę dzielącą sferę na dwie części: górną (reszty w otoczeniu grupy bocznej reszty zawierającej atom centralny) i dolną.

Hamelryck (2005) pokazał, że taka miara jest kompromisem pomiędzy parametrem zanurzenia reszty (Rozdział 2.3.6), który ma tendencję do opisu reszt jako znajdujących się w rdzeniu białka, a powierzchnią wyeksponowania do rozpuszczalnika (Rozdział 2.3.3) — z tendencją do opisu reszt jako wyeksponowanych.

# 3 DYNAMIKA BIAŁEK

*"Indeed the protein molecule model resulting from the X-ray crystallographic observations is a »platonic« protein, well removed in its perfection from the kicking and screaming »stochastic« molecule that we infer must exist in solution."*

— Weber, G., *Adv. Protein Chem., 1975, 29, 1-83*

W Rozdziale 1.1 przedstawiłem kilka przykładów zastosowań informacji o dynamice białek. W tym rozdziale opiszę doświadczalne i teoretyczne metody jej obserwacji.

## 3.1 METODY DOŚWIADCZALNE

Skupię się na najczęściej używanych metodach badających dynamikę białka na poziomie atomowym, pomimo że istnieją również inne interesujące techniki, jak na przykład metody badające dynamikę zmian kształtu i rozmiaru makromolekuły (SAXS (Doniach, 2001)) czy ruch białka w żywej komórce (FRET (Day i Schaufele, 2008), spektroskopia fluorescencyjna pojedynczej cząsteczki[1] (Fitter i in., 2011))[2].

### 3.1.1 Rentgenografia strukturalna

Technika ta jest najczęściej wykorzystywaną metodą przy wyznaczaniu struktury trzeciorzędowej białek (około 90% struktur zdeponowanych w bazie PDB rozwiązano z jej użyciem) i innych makromolekuł. Jakkolwiek, prócz wyznaczania struktury, eksperyment ten pozwala dostarczyć informację o dy-

1 Ang.: *Single molecule fluorescence spectroscopy*.
2 Rozdział ten napisałem opierając się głównie na pracy przeglądowej Boehr i in. (2006).

namice — zwykle o ruchach termicznych w krysztale, wyrażonych poprzez (an)izotropowy czynnik temperaturowy.

Izotropowy czynnik temperaturowy (B) (ang.: *B-factor*, czynnik Debye-Waller'a), wyraża fluktuacje atomów poprzez zależność:

$$B_i = \frac{8\pi^2 \left\langle (\Delta R)^2 \right\rangle_i}{3} \tag{3.1}$$

gdzie:

$$\left\langle (\Delta R)^2 \right\rangle_i = \frac{1}{T} \sum_{t_j=1}^{T} \left| \vec{x_i}(t_j) - \langle \vec{x_i} \rangle_T \right|^2 \tag{3.2}$$

w którym:

$t_j$ indeks konformacji ze zbioru T,
$\vec{x_i}$ wektor położenia atomu i.

B jest miarą nieoznaczoności położenia ciężkich atomów uzyskaną w trakcie rozwiązywania struktury rentgenograficznie (dopasowania obliczonych czynników struktury — zawierających czynnik temperaturowy — do obserwowanych w doświadczeniu czynników struktury).

Czynnik Debye-Waller'a opisuje ruchy termiczne atomów w krysztale, ale również nieuporządkowanie w poszczególnych komórkach elementarnych, defekty sieci krystalicznej czy szum powstały na etapie poprawiania modelu, co stwarza problemy z interpretacją tej wartości. Dodatkowo, czynnik ten nie zawiera informacji o skali czasowej ruchów termicznych (a więc posiadamy jedynie amplitudy wychyleń, brak jest natomiast częstości drgań) i nie koreluje zbyt dobrze z fluktuacjami z symulacji dynamiką molekularną (Tabela 1 w Pracy D, strona 115).

Pomimo wymienionych niedoskonałości, czynnik temperaturowy wykorzystuje się dość szeroko w badaniu dynamiki białek (Lu i in., 2006; Meinhold i Smith, 2005; Phillips, 1990) czy przy walidacji mechanicznych modeli dynamiki białek (typu ENM) (Bahar i Rader, 2005; Bahar i in., 1997; Kundu i in., 2002; Yang i in., 2009; Haliloglu i in., 1997).

*Rentgenografia z rozdzielczością czasową*

W pewnych specjalnych warunkach rentgenografia strukturalna pozwala uzyskać informację o zmianach w strukturze w funkcji czasu. Otrzymuje się wtedy poszczególne stany (etapy pośrednie) przebiegu procesu w pewnych odstępach czasowych (Moffat, 1998).

Mimo interesujących możliwości, technika ta nie jest uniwersalna — stosuje się ją jedynie dla szczególnych układów z użyciem wyspecjalizowanych urządzeń (Hajdu i in., 2000). Dodatkowo, by przeprowadzić tego typu eksperyment, układ musi spełniać kilka warunków:

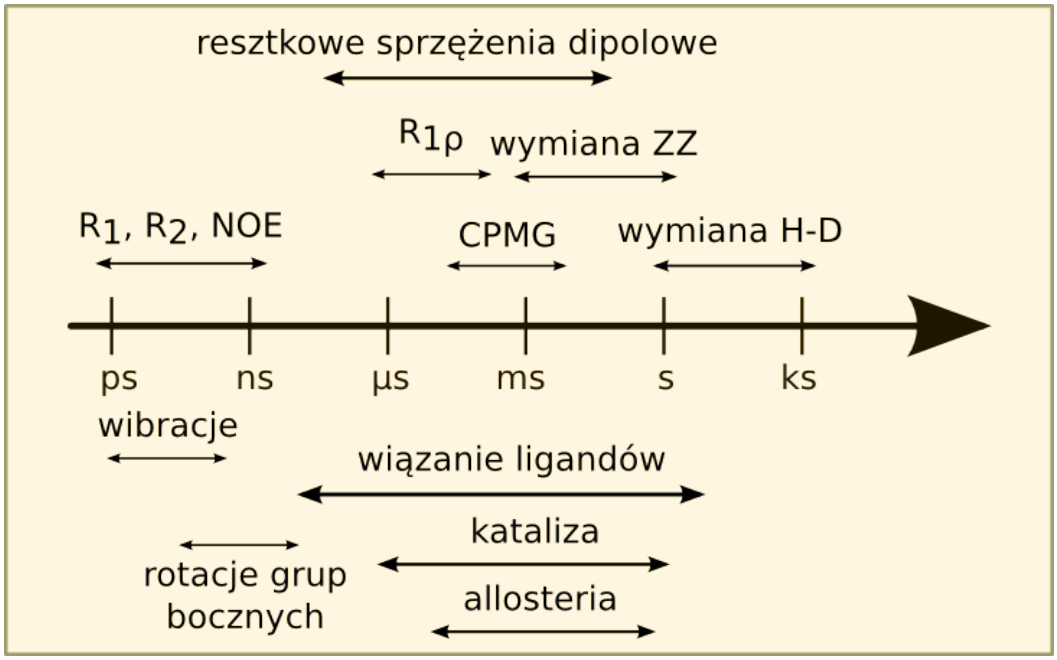
- makromolekuła musi być aktywna w krysztale,
- musi istnieć metoda wyzwalania reakcji, która nie narusza sieci krystalicznej i przebiega w całym krysztale jednolicie,
- stany pośrednie muszą być obecne w stosunkowo dużym stężeniu (powyżej 25%).

### 3.1.2 Spektroskopia jądrowego rezonansu magnetycznego, NMR

Tradycyjnie spektroskopię NMR wykorzystuje się przy wyznaczaniu struktur związków w roztworze, głównie małocząsteczkowych związków organicznych, ale również białek o średniej masie.

Spektroskopia NMR jest niewątpliwie najlepszą doświadczalną techniką analizy ruchu białek, pozwalającą na badanie ich w skalach czasowych odpowiadających istotnym przemianom biologicznym. Używając wysokorozdzielczej, wielowymiarowej spektroskopii NMR można uzyskać szczegółowe informacje na temat dynamiki na poziomie poszczególnych reszt aminokwasowych, czy atomów. Techniką tą można badać molekuły w obrębie skal czasowych od $10^{-12}$ s do $10^{5}$ s, co pokrywa wszystkie istotne zdarzenia dynamiki białek (Rysunek 6).

Dodatkowo, podejście TROSY pozwala badać duże układy (900 kDa kompleks czaperonu GroEL (Fiaux i in., 2002)), choć typowy eksperyment NMR wyznaczania struktury wykorzystuje się dla białek o masie poniżej 40 kDa (Homans, 2004).




**Rysunek 6:** Zakresy zdarzeń w dynamice białek oraz techniki NMR pozwalające na badanie danych zakresów. Rysunek utworzony na podstawie pracy Boehr i in. (2006).
CPMG: sekwencja pulsów Carr-Purcell-Meiboom-Gill; zz exchange (EXSY): technika mierząca wymianę informacji pomiędzy spinami w funkcji czasu pomiędzy głównymi, a pobocznymi sygnałami; wymiana H-D: technika opierająca się na obserwacji, że reakcje zachodzące szybko bądź w rdzeniu białka będą mniej podatne na podstawienie protu deuterem; $R_1$, $R_2$: pomiary czasów relaksacji podłużnej i poprzecznej; $R_{1\rho}$: pomiary czasu relaksacji w rotującym układzie współrzędnych.

*Skala czasowa pikosekund–nanosekund*

W tym zakresie czasowym zachodzą fluktuacje łańcucha głównego i grup bocznych. Bada się je mierząc trzy parametry: czas relaksacji podłużnej, czas relaksacji poprzecznej oraz NOE stanu stacjonarnego. Wartości te są następnie interpretowane w zakresie czasów korelacji ruchów i parametru $S^2$ (ang.: $S^2$ *order parameter*).

$S^2$ jest ogólną miarą ruchów kątowych wektorów łączących poszczególne jądra (Brüschweiler i Wright, 1994). Można ją zastosować do otrzymania wartości amplitud fluktuacji, po transformacji układu współrzędnych z kątowego na kartezjański (Haliloglu i Bahar, 1999).

Dodatkowo parametr ten można stosować przy szacowaniu wartości entropii konformacyjnej i jej wpływu na wiązanie substratu z enzymem (Wang i in., 2005; Frederick i in., 2007).

Większość doświadczeń w tej skali czasowej skupia się na wektorach wiązań N–H łańcucha głównego lub wybranych grup bocznych, choć uwzględniając dodatkowo wiązania $^{13}CO$–$^{13}C\alpha$ można dokładniej opisać dynamikę łańcucha głównego (Vugmeyster i Ostrovsky, 2011).

*Skala czasowa mikrosekund–milisekund*

W skali tej odbywa się wiele istotnych procesów, jak allosteria, wiązanie substratów, kataliza czy zwijanie łańcuchów białkowych.

W tym zakresie definiuje się parametr $R_{ex}$ – szybkość relaksacji wynikającej z wymiany pomiędzy konformacjami. Nie istnieje prosta zależność między ruchami w skali ps–ns, a µs–ms: w niektórych enzymach wiązanie liganda zmniejsza ruchy skali ps–ns, zwiększając — bądź nie zmieniając — ruchów skali µs–ms (Tabela 2 w Boehr i in. (2006)).

Najczęściej stosowaną metodą pomiarów w tej skali jest dyspersja relaksacji $R_2$ (ang.: $R_2$ *relaxation dispersion*) z pulsami CPMG i badanie jąder atomowych $^{15}N$ oraz $^{13}C$ (Mittermaier i Kay, 2006; Palmer i in., 2001).

Innym podejściem w tym zakresie czasowym jest RDC (Salmon i in., 2011; De Simone i in., 2011; Lindorff-Larsen i in., 2005), która jest o tyle interesująca, że za jej pomocą można równocześnie wyznaczyć strukturę i zbadać dynamikę.

## 3.2 METODY TEORETYCZNE

Hipoteza hiperpowierzchni energii potencjalnej zakłada, że stan natywny białka wyraża się przez minimum energii swobodnej Gibbsa, uwzględniającej entalpię i entropię układu. Potencjał ten przybliża się polami siłowymi, tj. modelami oddziaływań pomiędzy poszczególnymi atomami, gdzie parametry skalujące czy stałe siłowe zostały otrzymane z obliczeń mechaniki kwantowej lub obserwacji struktur otrzymanych z doświadczenia.

Zwykle taki potencjał składa się z sumy poszczególnych udziałów, np. dla pola siłowego AMBER (Cornell i in., 1995) definiuje się go następująco:

$$V(R) = \sum_{wiazania} k_a(x_i - x_{i,0})^2 + \sum_{katy} k_b(\theta - \theta_0)^2 + \\ + \sum_{\substack{katy \\ dwuscienne}} A_c\left[1 + \cos(n\omega)\right] + \sum_{i \neq j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \sum_{i \neq j} V_{Lennard-Jones}(r_{ij}) \quad (3.3)$$

gdzie zmiennymi są $x, r_{ij} = \|x_i - x_j\|, q, \omega, \theta$, pozostałe symbole wyrażają parametry; człon $V_{Lennard-Jones}(r_{ij})$ wyraża oddziaływania przyciągające/odpychające pomiędzy parami atomów.

Poprawnie skonstruowane pole siłowe powinno opisywać stan natywny poprzez minimum funkcji potencjału — minimum globalne na hiperpowierzchni energii potencjalnej.

Minimalizując funkcję $V(R)$ zwykle otrzymuje się poprawną geometrię danego układu (lub geometrię dla minimum lokalnego), chcąc jednak przeprowadzić symulację układu w ruchu, należy skorzystać z dynamiki molekularnej, bądź innych metod próbkowania przestrzeni konformacyjnej.

Ze względu na stosowane różne rozwiązania przy tworzeniu pól siłowych, wyniki symulacji tego samego układu z użyciem różnych pól siłowych prawdopodobnie będą się — bardziej, lub mniej — różniły (Rueda i in., 2007; Guvench i MacKerell, 2008).

### 3.2.1 Dynamika molekularna, metody deterministyczne[3]

Wykorzystując uprzednio wspomniane potencjały razem z równaniami ruchu Newtona o małym, rzędu femtosekund, kroku czasowym (gdzie siła $F(R) = -\nabla V(R)$, natomiast prędkości początkowe losowane są z rozkładu Maxwella-Boltzmanna dla danej temperatury) można przeprowadzić dokładne — o rozdzielczości poszczególnych atomów — symulacje zachowania molekuły białka w środowisku komórki.

Obecnie istnieje wiele zaawansowanych pakietów pozwalających na przeprowadzenie symulacji i analizę wyników dynamiki białek i innych makromolekuł, np.:

- AMBER (Case i in., 2005),
- CHARMM (Brooks i in., 2009),
- GROMACS (Hess i in., 2008),
- NAMD (Phillips i in., 2005),
- TINKER (Ponder i Richards, 1987)[4].

Dynamikę molekuarną uważa się za jedną z najlepszych metod symulacji — za jej pomocą wytłumaczono i zaproponowano wiele mechanizmów reakcji z udziałem białek, m.in.: procesy zwijania krótkich białek (Lindorff-Larsen i in., 2011), wiązania ligandów (Buch i in., 2011), efekty allosteryczne (Chiappori i in., 2012), zwijanie amyloidów (Lee i Ham, 2011), transport przez błonę komórkową (Johnston i Filizola, 2011).

Głównym problemem tego podejścia jest wysoki koszt obliczeniowy. Mając na uwadze konieczność użycia małego kroku czasowego ($10^{-15}$ s) przy całkowaniu równań Newtona, osiągnięcie biologicznego czasu rzędu 1 ms wymaga wykonania około $10^{12}$ kroków. Procesy biologiczne zachodzą w środowisku wodnym, przez co konieczne jest dodanie modeli wody do badanego systemu, co drastycznie zwiększa złożoność układu (około $10^5$ atomów uwzględniając cząsteczki rozpuszczalnika) i jednocześnie koszt obliczeń (Lane i in., 2012; Dror i in., 2012).

3 Choć dynamika molekularna również wykorzystuje metody stochastyczne (np. w równaniach Langevina), postanowiłem podzielić ten dział w ten sposób, że w niniejszym rozdziale opiszę metody, które uwzględniają czas *explicite*, w odróżnieniu od metod z Rozdziału 3.2.2, gdzie czas oszacować można jedynie zgrubnie.

4 Trzy ostatnie są pakietami darmowymi, dodatkowo GROMACS udostępniany jest na licencji wolnego oprogramowania i intensywnie rozwijany przez wolontariuszy.

Ostatnio poczyniono postępy w rozwoju sprzętu i oprogramowania, co umożliwiło badanie dłuższych czasów biologicznych i większych układów. Sprzętowymi udoskonaleniami ostatnich lat były:

- możliwość obliczeń zmiennoprzecinkowych na procesorach kart graficznych z wykorzystaniem technologii NVIDIA CUDA
  – pakiet ACEMD (Harvey i in., 2009) pozwala na osiągnięcie czasu 130 ns symulacji dziennie na jednym procesorze dla układu 23000 atomów[5],

- stworzenie komputera ANTON dedykowanego dynamice molekularnej (Chow i in., 2008a)[6]
  – ANTON pozwala na osiągnięcie czasu 471 ns dziennie dla układu 23000 atomów (Chow i in., 2008b),

- dedykowany obliczeniom dynamiki molekularnej komputer MDGRAPE-3 (Taiji i in., 2003).

Postęp w dziedzinie oprogramowania odbył się głównie za sprawą:

- rozwoju projektu BOINC (Anderson, 2004) umożliwiającego obliczenia rozproszone na komputerach wolontariuszy z wykorzystaniem CPU (*Folding@Home*) oraz GPU (GPUGRID),

- implementacji obliczeń z wykorzystaniem GPU w tradycyjnych pakietach MD (GROMACS, AMBER, NAMD),

- opracowania specjalnych algorytmów, np. REMD (Sugita i Okamoto, 1999), które przyspieszają zbieżność do minimum globalnego (w procesie zwijania białka).

Szacuje się, że wraz z rozwojem sprzętu i oprogramowania, około roku 2030 będzie możliwe przeprowadzenie symulacji pełno-atomowej dynamiki rybosomu w skali milisekund (Lane i in., 2012; Vendruscolo i Dobson, 2011).

Pomimo wielu sukcesów tej metody, osiągnięcie istotnych dla poznania funkcji białek mechanizmów (Rysunek 6 na stronie 22) jest wciąż zbyt kosztowne obliczeniowo, w związku z czym opracowuje się i intensywnie rozwija modele gruboziarniste (ang.: *coarse-grained*), cieszące się ostatnio coraz

5 Dla porównania: 16 procesorów E5-2670 wykona to zadanie z wydajnością 21 ns/dzień.
6 Dzięki tej technologii pomyślnie przeprowadzono rekordową symulację 1 milisekundy procesu zwijania 58. aminokwasowej aprotyniny wołowej (BPTI) (Shaw i in., 2010).

większym zainteresowaniem, oraz niedeterministyczne sposoby próbkowania hiperpowierzchni energii potencjalnej.

### 3.2.2 Metody stochastyczne, Monte Carlo

W odróżnieniu od metod dynamiki molekularnej, w podejściu Monte Carlo (MC) stosuje się zgoła inne rozwiązanie problemu eksploracji hiperpowierzchni energii potencjalnej.

Zamiast całkować równania ruchu małym krokiem czasowym, powierzchnia energii potencjalnej próbkowana jest losowo. Dzięki takiemu podejściu można znacznie zredukować koszt obliczeniowy przez dyskretyzację ruchów i dyskretyzację przestrzeni (czego nie można zrobić w przypadku MD, gdzie przestrzeń musi być ciągła). Dyskretyzacja ruchów jest wymagana, bowiem bez stosowania równań ruchu nieznane są siły poprzedniego kroku. Definiuje się więc zestawy ruchów lokalnych (perturbacji), naśladujące możliwe (fizyczne) ruchy białka, a wygenerowana konformacja akceptowana jest poprzez prawdopodobieństwo przejścia ze stanu poprzedniego do obecnego $P(R \to R')$.

Metropolis i in. (1953) zaproponowali takie kryterium (*asymetryczny schemat Metropolisa*) akceptacji przejścia pomiędzy stanami, uwzględniające rozkład Boltzmanna stanów zależny od temperatury:

$$P(R \to R') = \begin{cases} 1 & \text{gdy} \quad E' \leqslant E \\ e^{-\left(\frac{E'-E}{k_B T}\right)} & \text{gdy} \quad E' > E \end{cases} \tag{3.4}$$

gdzie:

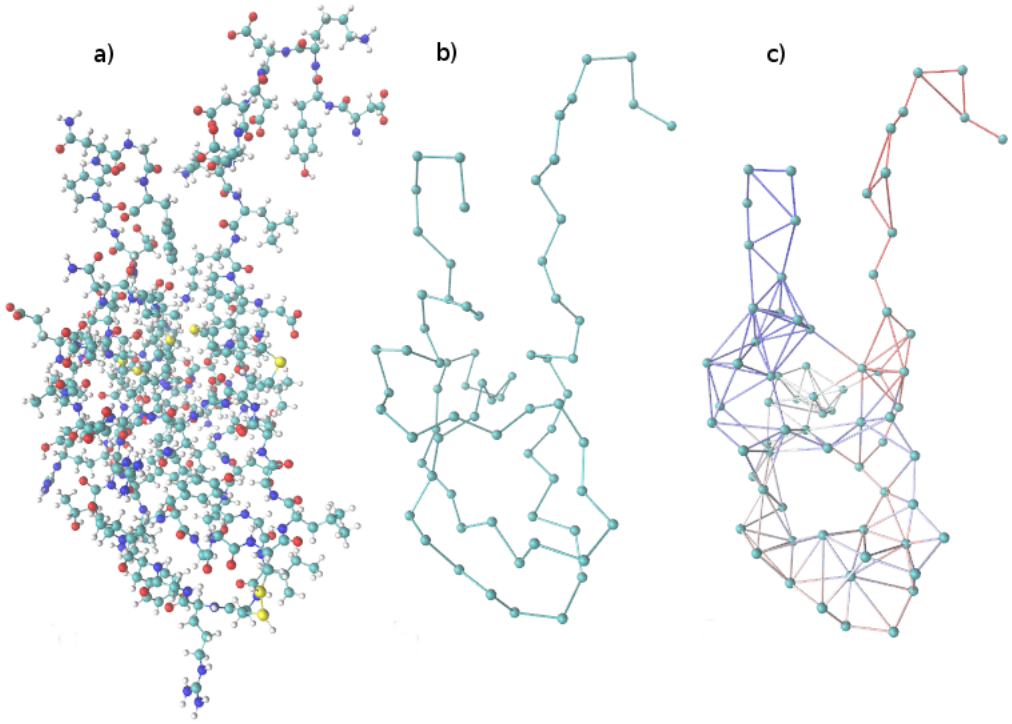
- $k_B$ stała Boltzmanna,
- $T$ temperatura,
- $E'$ energia stanu obecnego,
- $E$ energia stanu poprzedniego.

Zatem gdy energia nowego stanu jest mniejsza bądź równa energii stanu poprzedniego, nowy stan jest akceptowany — w pozostałych przypadkach akceptuje się nowy stan wtedy, gdy $e^{-\left(\frac{E'-E}{k_B T}\right)} \geqslant L$ ($L \in [0, 1]$, liczba losowa z rozkładu płaskiego).

Pomimo braku zmiennej czasowej, można wyobrazić sobie, że taka sekwencja stanów będzie zawierała konformacje opisujące ewolucję układu w czasie w danej temperaturze, np. przejścia pomiędzy dwoma minimami lokalnymi. Istotnie, tego typu podejście wykorzystali Kmiecik i Kolinski (2007) proponując ścieżki zwijania kilku krótkich białek globularnych, startując z całkowicie rozwiniętej losowej struktury.

# 4 MODELE GRUBOZIARNISTE

Modele gruboziarniste (zredukowane) tworzy się poprzez zastąpienie grupy atomów pseudo-atomem (centrum oddziaływania) reprezentującym właściwości danej grupy, bądź używanie tylko wybranych atomów (Rysunek 7). Takie uproszczenie redukuje liczbę stopni swobody, tym samym zmniejszając czas potrzebny na obliczenia oddziaływań.




**Rysunek 7:** Model gruboziarnisty na przykładzie struktury białka 1adz: a) reprezentacja pełno-atomowa; b) redukcja do poziomu atomów Cα; c) sieć oddziaływań w modelu ENM przy zastosowaniu promienia odcięcia 6 Å.

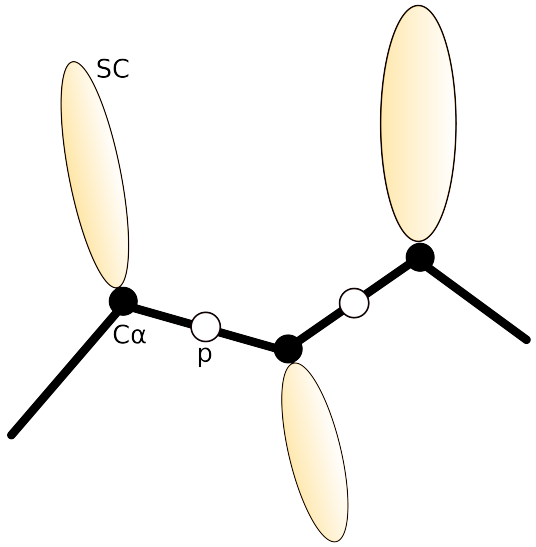
Stopień redukcji zależy od wielkości badanego układu — istnieją modele gdzie „ziarno" jest wielkości kilkudziesięciu reszt aminokwasowych (Lasker i in., 2012), z drugiej zaś strony najniższy stopień redukcji modelu pełno-

atomowego to pozbawienie go atomów wodoru[1]. Tak zredukowane modele są następnie opisywane specyficznym dla nich polem siłowym (Trylska, 2010).

Przykładami modeli gruboziarnistych mogą być:

MARTINI (Monticelli i in., 2008), w którym średnio na każde cztery atomy przypada jeden pseudo-atom (fragmenty zawierające pierścień zostały zredukowane jak 2:1).

UNRES (Liwo i in., 2005)[2], w którym zdefiniowane zostały centra oddziaływań grupy bocznej aminokwasowej (elipsoida) i pseudo-atomu pomiędzy dwoma atomami Cα — reprezentującego atomy wiązania peptydowego (Rysunek 8). Pole siłowe UNRES opiera się na potencjale średniej siły. Do symulacji ewolucji układu zaimplementowana została dynamika Langevina (uwzględniająca — za pomocą stochastycznych ruchów Browna — potencjał pochodzący od rozpuszczalnika); w algorytmie zastosowano również REMD, przyspieszającą zbieżność do minimum potencjału (globalnego).



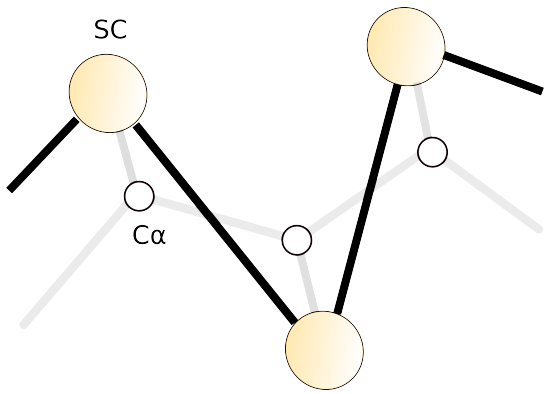

**Rysunek 8:** Model UNRES z zaznaczonymi pseudo-atomami (SC - elipsoida reprezentująca grupy boczne, p - pseudo-atom wiązania peptydowego) oraz atomami (Cα).

1 Tym samym można uznać, że modele rozwiązane z użyciem średniej rozdzielczości dyfraktometru są modelami zredukowanymi, nie posiadają bowiem współrzędnych położeń protonów.

2 Ang.: ***Un****ited* ***Res****idues*.

REDMD (Górecki i in., 2009) W modelu tym reprezentacja zredukowana jest do poziomu jednego pseudo-atomu na resztę aminokwasową. W pakiecie REDMD zastosowano pola siłowe w oparciu o ENM: zoptymalizowane dla rybosomu[3], proteazy HIV-1 oraz gruboziarnistego pola REACH (Moritsugu i Smith, 2007); przestrzeń konformacyjna próbkowana jest z użyciem dynamiki Brownowskiej.

SICHO (Kolinski i Skolnick, 1998)[4] Łańcuch białkowy reprezentowany jest przez pseudo-atomy w pozycji środka masy grup bocznych, uwzględniające atomy grup bocznych (Rysunek 9); próbkowanie przestrzeni konformacyjnej odbywa się z użyciem metod Monte Carlo (Rozdział 3.2.2) i pola siłowego opartego na statystyce znanych struktur białkowych.




**Rysunek 9:** Model SICHO. Kolorem żółtym zaznaczyłem pseudo-atomy grupy bocznej (SC). Kolorem białym zaznaczyłem atomy C$\alpha$, które są centrami oddziaływań, a ich pozycje obliczane są na podstawie położenia grup bocznych.

CABS (Kolinski, 2004b)[5] Łańcuch białkowy opisany jest poprzez atomy C$\alpha$, C$\beta$, pseudo-atom grupy bocznej oraz centrum oddziaływania położone pomiędzy kolejnymi atomami C$\alpha$ (Rysunek 11); próbkowanie odbywa się, jak w modelu SICHO, z użyciem metod Monte Carlo i statystycznego pola siłowego.

3 REDMD zawiera również gruboziarnisty model kwasów nukleinowych.
4 Ang.: ***Si**de **Ch**ain **O**nly*.
5 Ang.: ***C-a**lpha, C-**b**eta, **S**idechain*.

Ze względu na intensywne użycie tego modelu w trakcie badań, poświęciłem mu osobny rozdział (4.3), w którym bardziej szczegółowo opisałem jego założenia.

PRIMO (Gopal i in., 2010) — model ten opracowano redukując pełno-atomową reprezentację w taki sposób, by przejście od modelu zredukowanego do pełno-atomowego odbyło się bez straty dokładności (błędy na poziomie 0,1 Å) redukując strukturę średnio w stosunku 2:1. Obecnie nie użyto jeszcze tego modelu w przewidywaniu struktury czy modelowaniu dynamiki białek.

## 4.1 ANALIZA DRGAŃ NORMALNYCH I MODELE SIECI ELASTYCZNEJ (ENM)

Analiza drgań normalnych (NMA) przybliża powierzchnię energii potencjalnej (np. wyrażonej Równaniem 3.3) przez podział jej na niezależne drgania harmoniczne. Osiąga się to przez diagonalizację macierzy drugich pochodnych energii potencjalnej (a dokładniej macierzy $\mathbf{H} = \mathbf{M}^{-\frac{1}{2}}\ddot{\mathbf{V}}\mathbf{M}^{\frac{1}{2}}$, gdzie $\mathbf{M}$ to macierz mas), rozwiązując równanie $\mathbf{H}u_i = \omega_i^2 u_i$. Otrzymuje się wtedy 3N wartości własnych[6] ($\omega_i^2$, gdzie $\omega$ to częstość kołowa i-tego drgania) i odpowiadające im wektory własne, $u_i$. Przy założeniu, że $x = 0$ dla $t = 0$, można otrzymać wyrażenie na położenie atomów w funkcji czasu (Piela, 2005):

$$x_i(t) = x_i^0 + \frac{1}{\sqrt{m_i}} \sum_k^{3N} C_k u_{ik} \cos(2\pi\nu_k t + \phi_k) \tag{4.1}$$

gdzie:

6 Właściwie $3N - 6$ niezerowych wartości własnych.

$x_i(t)$ jedna ze współrzędnych atomu $i$ w czasie $t$,
$m_i$ masa $i$-tego atomu,
$C_k$ amplituda wychylenia, uwzględniająca temperaturę $C_k = \frac{\sqrt{2k_B T}}{2\pi\nu_k}$,
$u_{ik}$ współrzędna $k$ $i$-tego wektora własnego,
$\nu_k$ częstość drgań ($\omega = 2\pi\nu$),
$\phi_k$ faza.

Analiza drgań normalnych znajduje zastosowanie przy badaniu fluktuacji atomów białka, ruchów kolektywnych niskiej częstości, drgań termicznych i wynikających z tego możliwości, np. tworzenia zbiorów konformerów stanu natywnego (Kondrashov i in., 2007).

Tirion (1996) zaproponowała bardziej uproszczony model potencjału (model sieci elastycznej, ENM), gdzie atomy odległe od siebie nie dalej niż promień odcięcia ($R_c$) oddziałują ze sobą harmonicznie (Rysunek 10) za pomocą potencjału $V$, zdefiniowanego jako:

$$V = \gamma \frac{1}{2} \sum_{r^0_{i,j} < R_c} \left( r_{i,j} - r^0_{i,j} \right)^2 \tag{4.2}$$

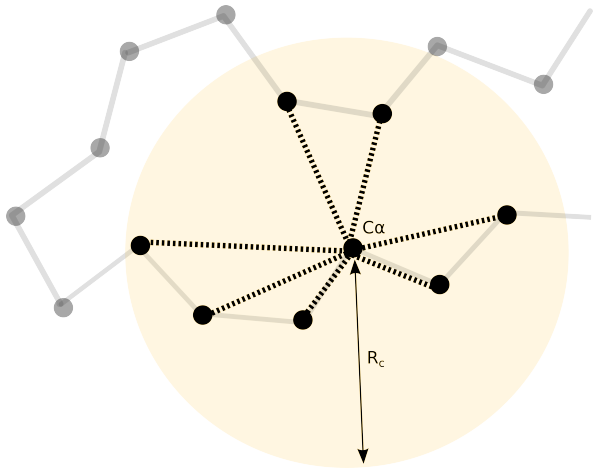
gdzie:

$\gamma$ stała siłowa sprężyny,
$r_{i,j}$ odległość pomiędzy atomem $i$-tym i $j$-tym,
$r^0_{i,j}$ odległość w strukturze początkowej,
$R_c$ promień odcięcia.

Macierz drugich pochodnych również się upraszcza, jej elementy można wyrazić w postaci (dla dwóch wymiarów):

$$h_{ij} = -\gamma \frac{(x_i - x_j)(y_i - y_j)}{\sqrt{m_i m_j} r^2_{ij}} \tag{4.3}$$

Okazuje się, że tak drastyczne uproszczenie potencjału z Równania 3.3 nie upośledza znacznie drgań normalnych o niskich częstościach. Bahar i in. (1997) zauważyli, że ze względu na rzadkość macierzy drugich pochodnych




**Rysunek 10:** Sieć oddziaływania pomiędzy atomem Cα a sąsiadującymi z nim atomami w odległości promienia odcięcia ($R_c$).

w ENM można ją bardziej uprościć, do modelu sieci Gaussowskiej (GNM), w której elementami macierzy **H** są:

$$h_{ij} = \begin{cases} -\gamma & \text{gdy} \quad i \neq j, r_{ij} \leqslant R_c \\ 0 & \text{gdy} \quad i \neq j, r_{ij} > R_c \\ -\sum_{i \neq j} h_{ij} & \text{gdy} \quad i = j \end{cases} \tag{4.4}$$

choć w wyniku takiego uproszczenia dostaje się jedynie informację o amplitudzie drgań (N wartości własnych) — traci się opis anizotropii fluktuacji (Sanejouand, 2013).

## 4.2 MODELE TYPU GŌ

Modele takie, opisane po raz pierwszy przeszło 40 lat temu (Taketomi i in., 1975), opierają się na założeniu, że stan natywny będzie opisany przez minimum energii potencjalnej, zaś potencjał zawierać będzie człon oddziaływań natywnych (wyrażonych przez mapę kontaktów stanu natywnego):

$$V(R) = V^{BB} + V^{S} + V^{N} \tag{4.5}$$

gdzie:

$V^{BB}$ człon opisujący potencjał harmoniczny pomiędzy atomami Cα w odległości równowagowej 3,8 Å,

$V^{S}$ potencjał narzucający więzy struktury drugorzędowej (np. w funkcji kątów $\Phi/\Psi$),

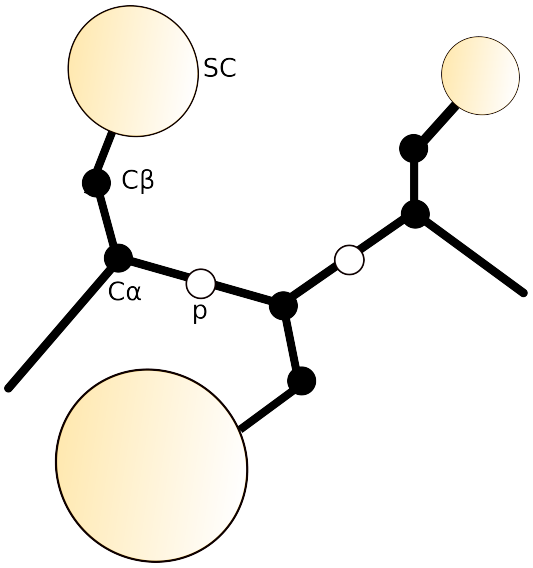
$V^{N}$ potencjał Lennarda-Jonesa dla oddziaływań natywnych.

W najprostszej postaci stosuje się reprezentację atomów Cα, oddziaływania natywne natomiast definiuje się z użyciem map kontaktów — oddziaływanie natywne to takie, w którym para atomów jest w bliższej odległości niż zadany promień odcięcia i nie sąsiaduje ze sobą wzdłuż sekwencji.

Mając tak zdefiniowany potencjał, wykorzystuje się np. równania ruchu Langevina do symulacji ewolucji układu (Sułkowska, 2007).

## 4.3 MODEL CABS

### 4.3.1 Reprezentacja łańcucha białkowego




**Rysunek 11:** Model CABS. Na rysunku zaznaczono pseudo-atomy reprezentujące grupy boczne (SC), środek wiązania peptydowego (p) oraz atomy Cα i Cβ.

Przestrzeń jest zdyskretyzowana za pomocą węzłów sieci prostej o gęstości 0,61 Å, na którą nanizane są atomy C$\alpha$. Wektory $v$, łączące kolejne atomy C$\alpha$ są w przestrzeni liczb całkowitych ($\mathbb{Z}$), a ich długość może być w zakresie $29 \leqslant \|v\|^2 \leqslant 49$ jednostek siatkowych (0,61 Å), co umożliwia fluktuacje odległości C$\alpha$-C$\alpha$ w zakresie 3,28-4,27 Å, wyrażone poprzez 800 wektorów. Dzięki temu można każdemu atomowi C$\alpha_i$ przyporządkować liczbę z zakresu 1-800 — odwołującą się do danego wektora — i na jej podstawie dostać położenie atomu C$\alpha_{i+1}$, a więc sekwencję N atomów C$\alpha$ można opisać N liczbami całkowitymi i współrzędnymi siatkowymi pierwszego atomu.

Posiadając z kolei współrzędne atomów C$\alpha_{i-1}$, C$\alpha_i$, C$\alpha_{i+1}$ można wyznaczyć położenie atomu C$\beta_i$ (którego współrzędne są już w przestrzeni liczb rzeczywistych, $\mathbb{R}$). Położenie grupy bocznej (również w $\mathbb{R}$) określa się na podstawie konformacji (kątów torsyjnych) atomów łańcucha głównego, wyróżniając dwa typy: *extended* oraz *compact*.

Dodatkowo, pomiędzy dwoma kolejnymi atomami C$\alpha$ zdefiniowane jest centrum oddziaływania odpowiadające za oddziaływania typu wiązań wodorowych łańcucha głównego.

Tak zredukowaną reprezentację przedstawiłem na Rysunku 11.

### 4.3.2 Próbkowanie przestrzeni konformacyjnej

W jednym kroku czasowym dynamiki MC następuje zmiana położeń losowo wybranych fragmentów łańcucha: dwa ruchy końca łańcucha, $10(N-2)$ ruchów dwóch wiązań C$\alpha$–C$\alpha$ (czyli jednego atomu), $N-3$ ruchów trzech wiązań, $N-24$ ruchów polegających na przesunięciu na małą odległość fragmentów łańcucha o długości 4-22 wiązań oraz $N-24$ ruchów „pełzających" wzdłuż łańcucha.

Nowy stan jest akceptowany z użyciem kryterium Metropolisa (Równanie 3.4), natomiast cała symulacja algorytmem CABS może być przeprowadzona w zależności od potrzeb:

- w warunkach izotermicznych — gdy w czasie trwania symulacji zadana jest stała temperatura. Podejście to zastosowałem przy symulacji dynamiki stanu natywnego białek (Praca E),

- z zastosowaniem symulowanego schładzania — gdy w trakcie symulacji temperatura układu (wpływająca na częstość akceptacji nowej konformacji) jest stopniowo obniżana, co w pewnym stopniu pozwala na wyeliminowanie problemu blokowania układu w jednym z minimów lokalnych,

- z wykorzystaniem metody wymiany replik Monte Carlo (REMC, Swendsen i Wang (1986)). Metodę tę, w połączeniu z symulowanym schładzaniem, wykorzystałem w pracach dotyczących przewidywania struktur białek (Praca B, A oraz C).

Ostatnie podejście znacznie przyspiesza osiągnięcie minimum globalnego układu, co ułatwia znalezienie struktury natywnej białka. W metodzie REMC stosuje się zbiór kilkudziesięciu konformacji (replik), symulowanych niezależnie w różnych temperaturach (zwykle $\Delta T = const.$) [7]. Raz na jakiś czas repliki są wymieniane (tj. danej replice przypisuje się inną temperaturę symulacji), przy zastosowaniu następującego kryterium:

$$P(R(T_i) \leftrightarrow R(T_j)) = \begin{cases} 1 & \text{gdy} \quad \Delta \leqslant 0 \\ e^{-\Delta} & \text{gdy} \quad \Delta > 0 \end{cases} \tag{4.6}$$

gdzie:

| | |
|---|---|
| $P$ | prawdopodobieństwo wymiany replik pomiędzy dwoma temperaturami. Gdy $\Delta > 0$ stosuje się tu — podobnie jak w przypadku tradycyjnego kryterium Metropolisa (Równanie 3.4) — przyrównanie do losowo wybranej liczby $L \in [0, 1]$, |
| $R(T_i)$ | konformacja o temperaturze $T_i$, |
| $\Delta$ | $= \left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)\left(E(T_i) - E(T_j)\right)$, |
| $E(T_j)$ | energia repliki o temperaturze $T_j$. |

7 Temperatury powinny być dobrane tak, by histogramy energii poszczególnych replik częściowo się pokrywały, umożliwiając ich wymianę.

### 4.3.3 Potencjał

W modelu CABS zastosowany został potencjał statystyczny (właściwie potencjał średniej siły[8]), wyprowadzony na podstawie rozwiązanych eksperymentalnie struktur białek z bazy PDB. Potencjał wyrażony jest przez sumę poszczególnych członów oddziaływań (opisanych szczegółowo w pracy Kolinski (2004b)):

$$V(R) = \varepsilon_{sr} V^{sr} + \varepsilon_{sr,seq} V^{sr}_{seq} + \varepsilon_{hb} V^{hb} + \varepsilon_{rep} V^{rep} + \varepsilon_{lr} V^{lr} + V^{lr}_{seq} \tag{4.7}$$

gdzie:

- $V^{sr}$ — niezależne od sekwencji oddziaływania bliskiego zasięgu,
- $V^{sr}_{seq}$ — zależne od sekwencji oddziaływania bliskiego zasięgu (wyprowadzone osobno dla par w odległości w sekwencji 1-3, 1-4 oraz 1-5),
- $V^{hb}$ — potencjał wiązań wodorowych, zależny od odległości między centrami oddziaływań (p) oraz kątem między wektorami prostopadłymi do wektora łączącego sąsiednie atomy C$\alpha$,
- $V^{rep}$ — oddziaływania odpychające,
- $V^{lr}$ — niezależne od sekwencji oddziaływania dalekiego zasięgu,
- $V^{lr}_{seq}$ — zależne od sekwencji oddziaływania dalekiego zasięgu, wyróżniające trzy typy względnego ułożenia grup bocznych (iloczyn wektorów C$\alpha$–SG) — równoległy, antyrównoległy oraz pośredni,
- $\varepsilon$ — czynniki skalujące.

W przeciwieństwie do potencjałów stosowanych w MD (zwykle opartych na przybliżeniach mechaniki kwantowej dla małych układów), model CABS wykorzystuje potencjały statystyczne, wyprowadzone z bazy danych struktur białkowych w środowisku wodnym. Tego typu potencjały opierają się na założeniu, że częstość występowania danego stanu w bazie danych struktur

8 Warto zauważyć, że potencjał taki będzie uwzględniał oddziaływania z rozpuszczalnikiem.

białkowych będzie w przybliżeniu zgodna z rozkładem Boltzmanna (Sippl, 1995; Finkelstein i in., 1995):

$$\frac{N_i}{N_j} = e^{\frac{-(E_i - E_j)}{k_B T}} \tag{4.8}$$

gdzie:

$N_i$ liczba obsadzonych stanów o energii $E_i$.

Tak więc różnica energii dwóch stanów będzie równa:

$$E_i - E_j = -k_B T \ln\left(\frac{N_i}{N_j}\right) \tag{4.9}$$

By otrzymać wartość bezwzględną energii danego stanu, należy przyjąć pewien układ odniesienia, który będzie wyrażał energię wszystkich stanów. Jest to kluczowe — poprawność danego potencjału średniej siły w dużej mierze zależy od przyjętego układu odniesienia. Mając układ odniesienia, dostaje się energię bezwzględną:

$$E(R) = -k_B T \ln\left(\frac{N(R)}{\sum N(R')}\right) = -k_B T \ln(f(R)) \tag{4.10}$$

gdzie:

$R$ zmienna pomiędzy oddziałującymi ze sobą elementami (np. odległość między atomami, wartość kąta torsyjnego czy typ aminokwasu),

$N(R')$ liczba wszystkich stanów (układ odniesienia),

$f(R)$ częstość występowania w bazie danych stanu opisanego zmienną $R$.

# Część III

# Streszczenie prac stanowiących podstawę rozprawy

# 5 MODELOWANIE PĘTLI W STRUKTURACH BIAŁEK

## 5.1 WPROWADZENIE

Rozdział dotyczy Pracy A (Jamroz i Kolinski, 2010).

Kluczowymi elementami modelowania porównawczego białek są: modelowanie na podstawie dopasowania sekwencji białka-celu do sekwencji białka-szablonu oraz modelowanie fragmentów pętli (fragmentów niedopasowanych). Konformacje krótkich pętli można przewidzieć z dużą dokładnością przez dopasowanie fragmentów struktury z innych — niekoniecznie homologicznych — białek, bądź z użyciem metod poszukiwania minimum funkcji potencjału.

Dłuższe pętle udaje się pomyślnie modelować z zastosowaniem podejścia wieloskalowego, wykorzystującego gruboziarniste modelowanie *de novo* — w tym przypadku model CABS.

## 5.2 STRESZCZENIE PRACY

Mając zestaw testowy (ang.: *benchmark*) fragmentów pętli, przedstawiony w pracy Rossi i in. (2007) i zawierający różnorodne strukturalnie modele białek, rozszerzyłem go o pętle dłuższe — do 25. aminokwasów — stosując metodę DSSP (Rozdział 2.3.2, Kabsch i Sander (1983)) na obecnych w zestawie strukturach, a następnie wybierając fragmenty nie posiadające przypisanej struktury drugorzędowej, otrzymując całkowitą liczbę 186. fragmentów. Tak wybrane regiony przedstawiłem w Tabeli 1 Pracy A (strona 89).

Zestaw ten następnie wykorzystałem przy modelowaniu pętli z użyciem trzech metod: MODELLER, ROSETTA oraz CABS.

Modelowanie algorytmem MODELLER odbywało się z zastosowanie klasy `loop`. Przy jej użyciu zostało zaproponowanych 500 alternatywnych struktur, z których każda została oceniona potencjałem statystycznym DOPE (Shen i

Sali, 2006), wybierając model *top*, dla którego wartość potencjału była najniższa. Z otrzymanego zestawu wybrałem również model *best*, który był modelem najbliższym strukturze eksperymentalnej pod względem RMSD (Dodatek G.3) fragmentu pętli po optymalnym nałożeniu całości struktury na strukturę odniesienia.

Modelowanie algorytmem ROSETTA odbywało się z zastosowaniem aplikacji `loopmodel.*`[1] i wykorzystaniu metody *Cyclic Coordinate Descent* (Canutescu i Dunbrack, 2003; Wang i in., 2007) tworzenia konformacji pętli. W tym przypadku również stworzyłem 500 modeli, spośród których model *top* był najlepiej ocenionym modelem według potencjału ROSETTA.

W przypadku algorytmu CABS przeprowadziłem symulację REMC w warunkach symulowanego schładzania (temperatura: $2.0 \rightarrow 1.0$) z zadanymi więzami odległości na całą strukturę z wyłączeniem atomów fragmentu pętli. Długość symulacji została zadana tak, by otrzymać 200 modeli. Spośród tak stworzonych konformacji wybrałem model *top* jako medoid (model najbliższy — porównany miarą RMSD — średniej z trajektorii) całego zbioru[2].

W kolejnym podejściu 10 najlepszych (*top*) modeli otrzymanych z programu MODELLER użyłem jako struktury białek-szablonów w modelowaniu algorytmem CABS. Podobne podejście zastosowałem przy modelowaniu ROSETTA/CABS, uzyskując wyniki zbliżone do MODELLER/CABS. Jako że algorytm MODELLER generuje wynik znacznie szybciej niż ROSETTA, w pracy została opisana wyłącznie metoda MODELLER/CABS (CABS-*hybrid*).

1 Przykładowe polecenie uruchamiające aplikację:
`loopmodel.linuxgccrelease -database rosetta3_database -in::file::fullatom -loops::input_pdb 135l_18-27T -loops::loop_file 135lA.loop_file -loops::frag_files aa135lA09_05.200_v1_3 aa135lA03_05.200_v1_3 none -nstruct 500 -loops::build_initial -loops::ccd_closure -loops::random_loop -out::prefix LOOP -in::file::psipred_ss2 135lA.psipred_ss2 -mute core.io.database`

2 Początkowo stosowałem metodę analizy skupień K-średnich, jednak średnio dawała ona podobne wyniki, tj. medoid największego skupienia był praktycznie taki sam jak medoid z całego zbioru. Próbowałem również stosować potencjały opisane w Feng i in. (2010), lecz wyniki nie przedstawiały zauważalnej przewagi nad metodą wyboru medoidu.

## 5.3 WYNIKI I WNIOSKI

Wyniki przedstawiłem na Rysunkach 1–4 oraz Tabeli 2 Pracy A (strony 90–92).

Przy użyciu algorytmu CABS możliwe było wymodelowanie długich (16–25) pętli nieznacznie lepiej niż przy użyciu pozostałych algorytmów — średni RMSD: 8,11 Å w porównaniu z 8,39 Å (MODELLER) i 10,02 Å (ROSETTA). Użycie metody hybrydowej, tj. połączenia algorytmu CABS i MODELLER pozwoliło poprawić średnią wartość RMSD — z 8,11 Å do 7,87 Å w tym zakresie długości pętli. Metoda hybrydowa pozwoliła również na poprawę wyniku dla pętli o długości 7–12 aminokwasów (z 3,83 Å (CABS) do 2,23 Å (CABS-*hybrid*)).

Analizując średni RMSD modeli oznaczonych jako *best* można wysnuć wniosek, że żaden z algorytmów nie posiada dobrej metody wyboru najlepszego (najbliższego strukturze eksperymentalnej) modelu. Istotnie, jest to wciąż nierozwiązany problem, wynikający z uproszczeń stosowanych przy konstrukcji potencjałów.

Warto zauważyć, że uzyskane w ten sposób pseudo-trajektorie z modelowania wieloskalowego (po odbudowie modelu do reprezentacji pełno-atomowej) mogą być użyte w badaniu dynamiki pętli.

# 6 OPRACOWANIE PÓŁAUTOMATYCZNEJ METODY PRZEWIDYWANIA STRUKTUR BIAŁEK

## 6.1 WPROWADZENIE

Rozdział dotyczy Pracy B (Kmiecik i in., 2008) oraz Pracy C (Blaszczyk i in., 2012).

Proces modelowania struktur białek na podstawie samej sekwencji wydaje się procesem żmudnym, często wymagającym przeprowadzania za każdym razem tych samych procedur postępowania. Udostępnienie narzędzia wykorzystującego półautomatyczny protokół modelowania umożliwi zastosowanie komputerowych metod przewidywania struktury białek społeczności naukowej działającej w innych dziedzinach, pośrednio związanych z proteomiką.

Metodę określam jako półautomatyczną, bowiem etap dopasowania sekwencji — jako najbardziej istotny etap modelowania struktury — pozostawia się użytkownikowi.

## 6.2 STRESZCZENIE PRAC

W ostatnim czasie pokazano, że podejście *meta*-serwerów (serwerów wykorzystujących równocześnie wiele podejść dopasowania sekwencji celu do struktury białka-szablonu) pozwala na stworzenie konsensusowego dopasowania sekwencji, średnio lepszego od wyniku z poszczególnych metod osobno (Lundström i in., 2001; Bujnicki i in., 2001; Wallner i Elofsson, 2005; Kurowski i Bujnicki, 2003). Tak otrzymane dopasowanie sekwencji powinno zostać sprawdzone przez specjalistę i ewentualnie zmodyfikowane. Dopiero po

tym kroku można przejść do kolejnych etapów modelowania, zastosowanych w opisywanej tu procedurze.

Procedura umożliwia również modelowanie *de novo*, tj. z wykorzystaniem *jedynie* informacji o sekwencji białka-celu. Należy jednak mieć na uwadze, że modelowanie *de novo* nie sprawdza się — żadną ze znanych dziś metod — dla białek o długości łańcucha ponad ok. 120 aminokwasów.

Przepływ danych i zastosowane kroki modelowania w opracowanej procedurze przedstawiłem na Rysunku 12. Informacją przekazaną do algorytmu CABS jest sekwencja, struktura drugorzędowa (zredukowana do oznaczeń $\alpha$-helis oraz $\beta$-wstęg) oraz prostokątny potencjał kary, pochodzący z więzów odległości na atomy C$\alpha$, zdefiniowany jako:

$$V_{wiezy} = K \sum_i w_i \Delta_i \tag{6.1}$$

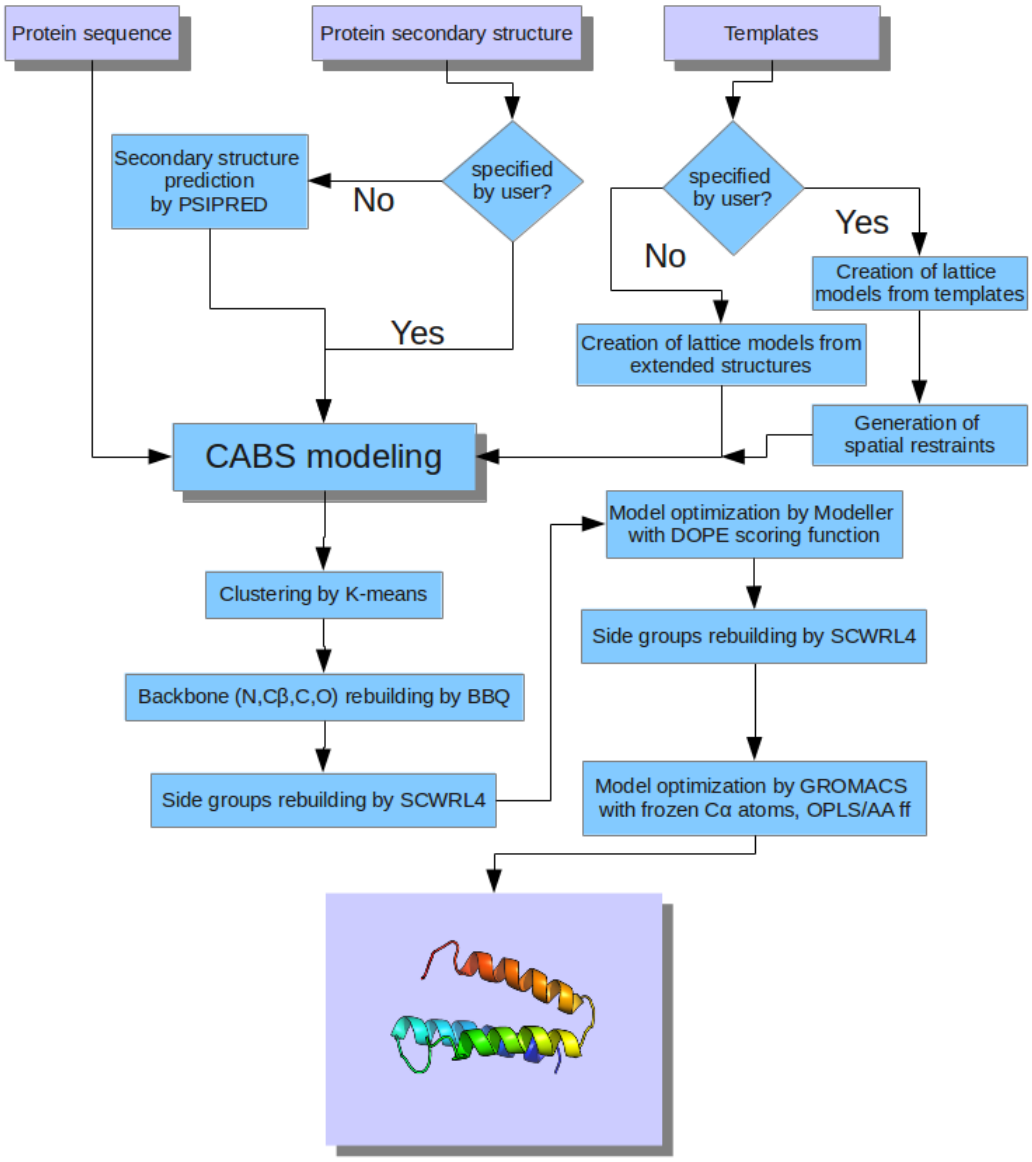
gdzie:

$$\Delta_i = \begin{cases} \overline{d_i} - \sigma_i - d_i & \text{gdy} \quad d_i < (\overline{d_i} - \sigma_i) \\ d_i - \overline{d_i} - \sigma_i & \text{gdy} \quad d_i > (\overline{d_i} + \sigma_i) \\ 0 & \text{gdy} \quad \overline{d_i} - \sigma_i \leqslant d_i \leqslant \overline{d_i} + \sigma_i \end{cases} \tag{6.2}$$

oraz:

- $d_i$ — odległość pomiędzy i-tą parą atomów w trakcie symulacji,
- $\overline{d_i}$ — średnia odległość pomiędzy i-tą parą atomów w zestawie struktur białek-szablonów,
- $\sigma_i$ — odchylenie standardowe i-tego więzu odległości,
- $w_i$ — częstość występowania i-tego więzu (równe 1 wtedy, i tylko wtedy, gdy para atomów, wokół której zdefiniowano więz występuje we wszystkich białkach-szablonach, tj. brak przerw w podpisaniu sekwencji dla tych atomów),
- $K$ — stała skalująca.

By zwiększyć swobodę zmian konformacyjnych łańcucha związanego powyższym potencjałem, pary i-te więzów były brane w odstępach: $(j, j+6)$, $(j, j+15)$ oraz $(j, j + \sum_l \lfloor k_l * 1.4 \rfloor)$ dla $k \in [14, 28]$ oraz $j \in [1, N]$ (oczywiście w zakresie długości łańcucha, N).




**Rysunek 12:** Przepływ danych i zastosowane kroki modelowania w opracowanej procedurze przewidywania struktury białek na podstawie sekwencji (i struktur białek-szablonów).

## 6.3 WYNIKI I WNIOSKI

Metodyka została, m.in. przeze mnie, intensywnie testowana podczas eksperymentu CASP9 (grupy *bujnicki-kolinski* oraz *LTB*[1], zajmując — kolejno — miejsca 26. i 28. spośród 174.). W rankingu liczby najlepiej przewidzianych modeli (Tablica 1) nasze grupy zajęły miejsca: 2. (*LTB*) oraz 4. (*bujnicki-kolinski*).

Początkowo opracowana procedura (opisana w Pracy B) została komercyjnie zastosowana w internetowej aplikacji *Protein Modeling Platform* spółki SELVITA[2].

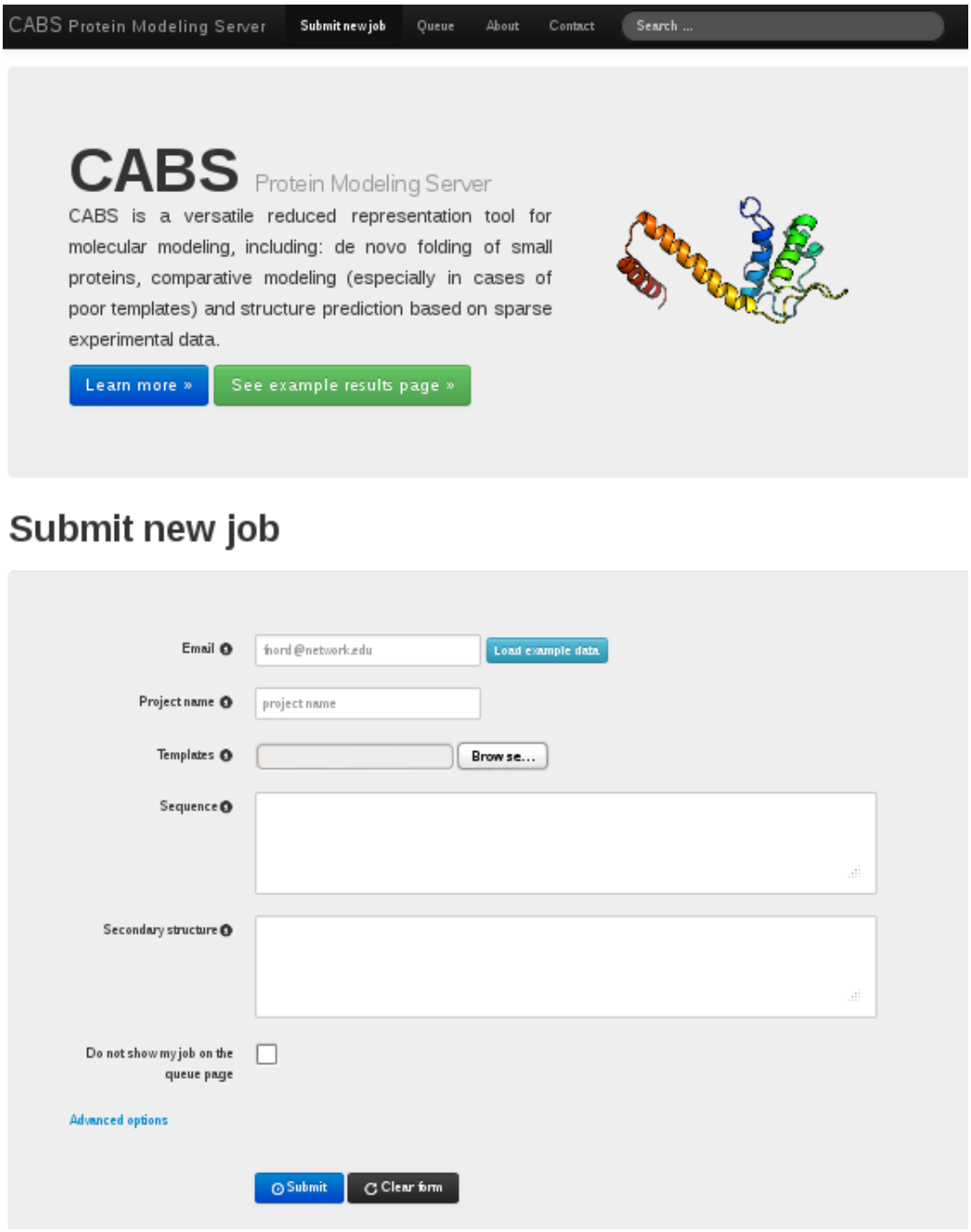
Aplikacja wykorzystująca metodykę z Rysunku 12 jest dostępna dla użytkowników pod adresem http://biocomp.chem.uw.edu.pl/CABSserver (zrzuty ekranu przedstawiłem na Rysunkach 14–16).

1 W grupie tej stosowaliśmy głównie struktury białek-szablonów uzyskane z serwerów uczestniczących w CASP9, wyniki zostały przedstawione w Pracy C (Rysunki 2 i 3 na stronie 107 przedstawiają miarę GDT_TS opisaną w Dodatku G.4).

2 Z wykorzystaniem metody BLAST przy wyszukiwaniu struktur, metody odbudowy grup bocznych opracowanej przeze mnie w Jamroz (2008), pominięciu kroku optymalizacji algorytmem MODELLER oraz z zastosowaniem znacznie bardziej zaawansowanego interfejsu użytkownika.

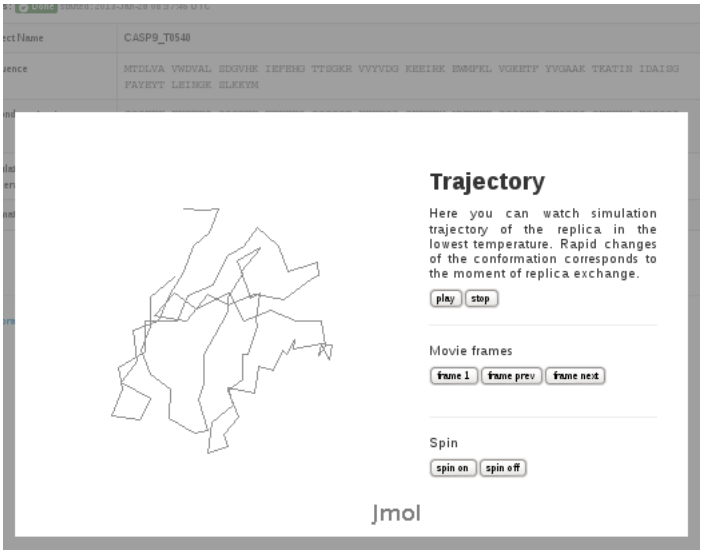
| Miejsce | Liczba najlepszych modeli | Nazwa grupy | Uczestnicy |
|---|---|---|---|
| 1 | 7 | PRMLS | Jimin Pei |
| 2 | 5 | LTB | Maciej Błaszczyk, Michał Jamróz, Andrzej Koliński |
| 2 | 5 | BAKER | James Thompson, Ray Wang, Firas Khatib, Michael Tyka, TJ Brunette, Dominik Gront, Frank DiMaio |
| 3 | 4 | ZHANG_AB_INITIO | Yang Zhang, Dong Xu |
| 4 | 3 | FEIG | Michael Feig, Kanagasabai Vadivel |
| 4 | 3 | MUFOLD | Dong Xu, Jingfen Zhang, Qingguo Wang, Yi Shang, Jiong Zhang, Zhiquan He, Yang Xu, Ioan Kosztin, Chao Zhang |
| 4 | 3 | BUJNICKI-KOLINSKI | Michał Jamróz, Maciej Błaszczyk, Janusz Bujnicki, Andrzej Koliński, Mateusz Warkocki, Kasia Mikołajczak |
| 4 | 3 | JONES-UCL | David Jones, Daniel Buchan, Domenico Cozzetto, Sean Ward |
| 5 | 2 | ELOFSSON | Arne Elofsson, Marcin Skwark, Bjorn Wallner, Arjun Ray |
| 6 | 2 | SWA_TEST | Kyle Beauchamp, Rhiju Das |

**Tablica 1:** Ranking według liczby modeli ocenionych jako najbardziej poprawne spośród wszystkich modeli przesłanych w trakcie trwania eksperymentu CASP9.

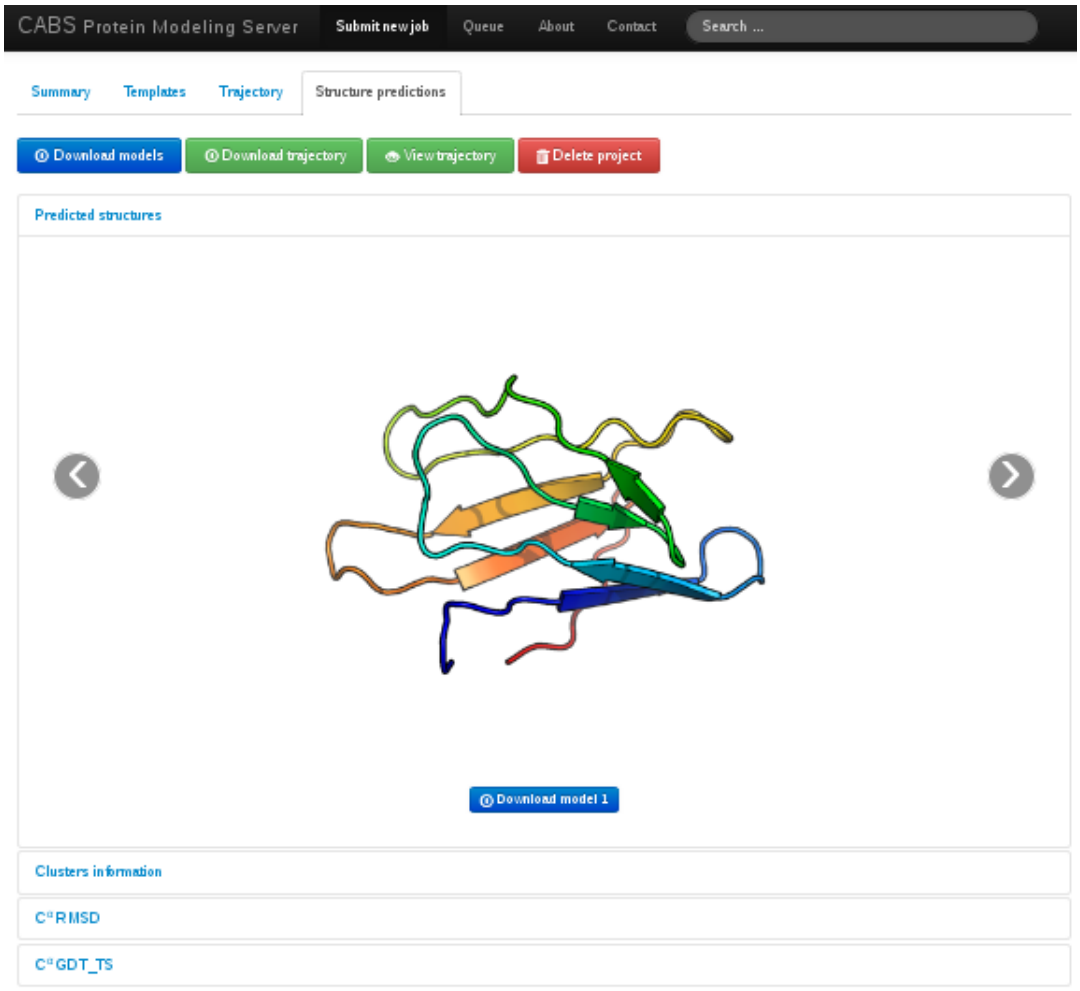



**Rysunek 13:** Interfejs wprowadzania danych użytkownika (sekwencja, struktury-szablony w formacie plików PDB, predykcja struktury drugorzędowej). Dodatkowo użytkownik może wybrać zakres temperatury symulacji oraz zmodyfikować (bądź dodać nowe) więzy odległości (*Advanced options*). Podając jedynie sekwencję aminokwasową uruchomi się procedura modelowania *de novo* z zastosowaniem metody PSIPRED (Jones, 1999) przewidywania struktury drugorzędowej.


**Rysunek 14:** W trakcie symulacji na bieżąco wyświetlane są podstawowe parametry łańcucha, takie jak promień żyracji, odległość między końcami oraz energia CABS.




**Rysunek 15:** Po zakończeniu symulacji, użytkownik może m.in. wyświetlić „film" z jej przebiegu (pseudo-trajektorię atomów Cα).




**Rysunek 16:** Aplikacja umożliwia pobranie pełno-atomowych modeli struktur reprezentatywnych (medoidów skupisk po zastosowaniu analizy skupień K-średnich).

# 7 OPRACOWANIE METODY PRZEWIDYWANIA WARTOŚCI FLUKTUACJI ATOMÓW W BIAŁKACH

## 7.1 WPROWADZENIE

Rozdział dotyczy Pracy D (Jamroz i in., 2012).

Obecnie biologia molekularna coraz bardziej skupia się na badaniu — metodami doświadczalnymi i teoretycznymi — dynamiki białek. Dzieje się tak, bowiem poznanie dynamiki białek pozwala na zrozumienie mechanizmów funkcjonowania białek w wielu procesach zachodzących w żywych komórkach (Teilum i in., 2009): w szczególności wiązania ligandów, reakcji enzymatycznych czy oddziaływań białko-białko.

Zrozumienie ruchliwości białek pozwoli na rozwinięcie metod komputerowego projektowania leków czy enzymów (Mandell i Kortemme, 2009; Lassila, 2010; Lill, 2011).

## 7.2 STRESZCZENIE PRACY

W pracy zbadałem zależność różnych opisów struktury białka na jego dynamikę, wyrażoną przez fluktuacje atomów C$\alpha$ w odniesieniu do struktury wyznaczonej doświadczalnie na dużym zestawie (głównie 10 ns) trajektorii MD, zawierającym 592 nieredundantne sekwencyjnie białka wszystkich głównych klas CATH (Orengo i in., 1997).

Zastosowałem parametry opisane w Rozdziałach 2.3.1–2.3.8 (strony 12–17), czynnik temperaturowy z eksperymentu rentgenografii strukturalnej (Roz-

dział 3.1.1, strona 20) oraz wartości fluktuacji otrzymane z gruboziarnistej metody GNM (Rozdział 4.1, strona 33):

$$\left\langle (\Delta R)^2 \right\rangle_i = k\mathbf{H}_{ii}^{-1}, \tag{7.1}$$

z zastosowaniem dwóch różnych konstrukcji macierzy Kirchhoffa (**H**): tradycyjnej, zaproponowanej przez Bahar (Równanie 4.4, przyjmując promień odcięcia, $R_c = 16\text{Å}$[1]) oraz modyfikację zaproponowaną przez Yang i in. (2009), gdzie elementami macierzy **H** są:

$$h_{ij} = \begin{cases} r_{ij}^{-2} & \text{gdy} \quad i \neq j \\ -\sum_{i \neq j} h_{ij} & \text{gdy} \quad i = j \end{cases} \tag{7.2}$$

Po wyznaczeniu parametrów mających największy wpływ na fluktuacje atomów, stworzyłem modele SVR zawierające różne ich kombinacje, starając się zmaksymalizować współczynnik korelacji Pearsona i zminimalizować średnie kwadratowe odchylenie wartości fluktuacji otrzymanych z modelu i fluktuacji pochodzących z dynamiki molekularnej.

## 7.3 WYNIKI I WNIOSKI

W pracy zaproponowałem szybką metodę przewidywania wartości fluktuacji na podstawie samej struktury białka; metoda pozwala otrzymać wyższą korelację w porównaniu z GNM (odpowiednio $r_{SVR,MD} = 0,669$, $r_{GNM,MD} = 0,646$) i — co istotniejsze — wynik otrzymywany jest w postaci wartości bezwzględnych[2], co przedstawiłem w Tabelach I i II Pracy D (strony 115–116).

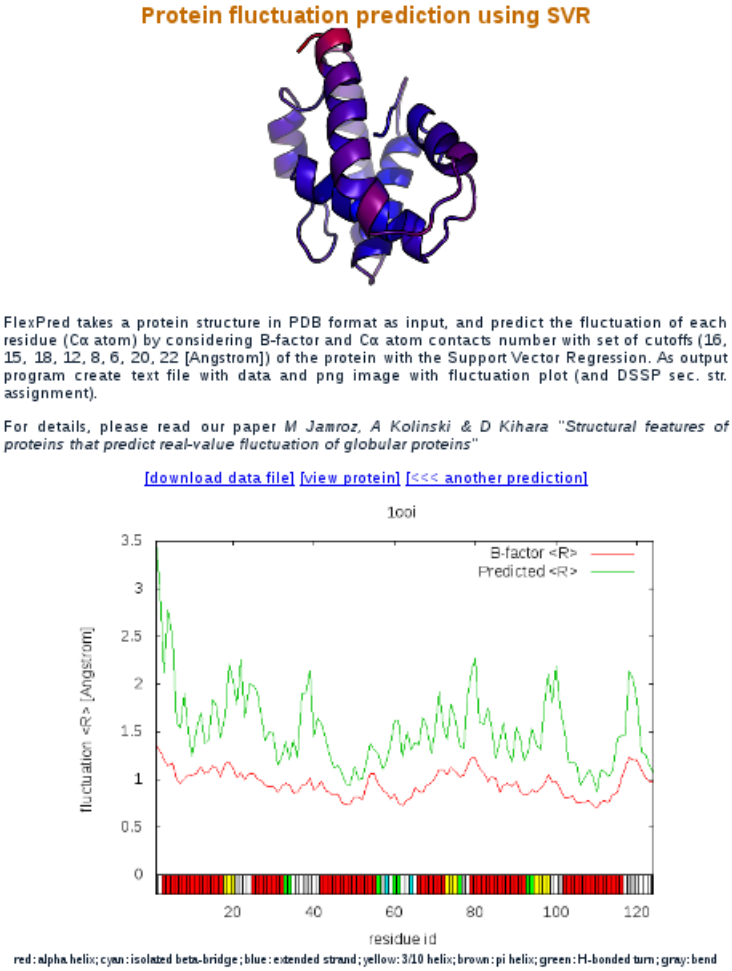
Otwarty kod programu wykorzystujący opisaną metodę, jak również interfejs www umożliwiający obliczenia (po zadaniu kodu białka z bazy PDB) na zdalnym serwerze jest dostępny poprzez stronę http://kiharalab.org/flexPred

1 Wartość ta została dobrana maksymalizując średni współczynnik korelacji Pearsona (Dodatek G.1) pomiędzy wartościami fluktuacji otrzymanymi z GNM z fluktuacjami z MD.

2 W przypadku GNM otrzymane wartości fluktuacji należy przeskalować nieznaną *a priori* stałą k (Równanie 7.1).

dla każdego zainteresowanego użytkownika (przykładowy wynik działania serwera pokazałem na Rysunku 17).

Jest to prawdopodobnie pierwsza do tej pory metoda pozwalająca na przewidywanie wartości fluktuacji z eksperymentu dynamiki molekularnej w czasie poniżej minuty. Dodatkowo, prawdopodobnie jako pierwsi przeanalizowaliśmy wpływ parametrów struktury na fluktuacje atomów w trakcie doświadczenia MD na tak dużym zestawie danych.




**Rysunek 17:** Wynik predykcji wartości fluktuacji atomów Cα dla białka o kodzie PDB 1ooi. W górnej części ekranu wyświetlana jest struktura pokolorowana względem wartości przewidzianych fluktuacji, kolorem zielonym na wykresie przedstawione są przewidziane wartości fluktuacji, zaś kolorem czerwonym — fluktuacje po przeliczeniu czynnika temperatury.

# 8 PORÓWNANIE DYNAMIKI GRUBOZIARNISTEGO MODELU CABS Z PEŁNO-ATOMOWYMI MODELAMI DYNAMIKI MOLEKULARNEJ

## 8.1 WPROWADZENIE

Rozdział dotyczy Pracy E (Jamroz i in., 2013).

Ze względu na koszt obliczeń, wykorzystanie pełno-atomowych pól siłowych i dynamiki molekularnej przy badaniu procesów zachodzących z udziałem białek jest ograniczone do relatywnie krótkich czasów biologicznych — w związku z tym zastosowanie modeli gruboziarnistych i próbkowania stochastycznego wydaje się atrakcyjną alternatywą.

## 8.2 STRESZCZENIE PRACY

Rueda i in. (2007) porównali cztery popularne pola siłowe stosowane w dynamice molekularnej (GROMOS, AMBER, OPLS/AA, CHARMM) przeprowadzając 10 ns symulacje z rozpuszczalnikiem w formie jawnej trzydziestu białek należących do różnych klas CATH. Z otrzymanych wyników wyciągnęli wniosek, że wszystkie cztery pola siłowe dają konsystentny obraz zachowania się łańcucha białkowego.

Dzięki uprzejmości autorów, otrzymaliśmy trajektorie dla 22. białek[1], dla których przeprowadziłem symulacje gruboziarnistym modelem CABS.

Starałem się tak zoptymalizować parametry potencjału CABS i czasu trwania symulacji, by zmaksymalizować średni współczynnik korelacji wartości

1 Pozostałe trajektorie zostały utracone w wyniku awarii dysków autorów pracy.

fluktuacji, przy czym zastosowałem walidację 2-krotną. Po otrzymaniu zadowalających wyników porównałem trajektorie otrzymane z algorytmu CABS z trajektoriami MD, stosując miary podobieństwa konformacji łańcucha ($\Omega$, Równanie 2, strona 127) oraz miary podobieństwa ruchów ($s(A,B)$, $\gamma_{AB}$, Równania — odpowiednio — 6 i 7, strona 127).

Dodatkowo, by porównać typ dyfuzji w poszczególnych trajektoriach oraz oszacować czas CPU potrzebny do wychylenia atomów z położeń na tę samą odległość, sporządziłem wykres funkcji autokorelacyjnej (Rysunek 5, strona 130).

## 8.3 WYNIKI I WNIOSKI

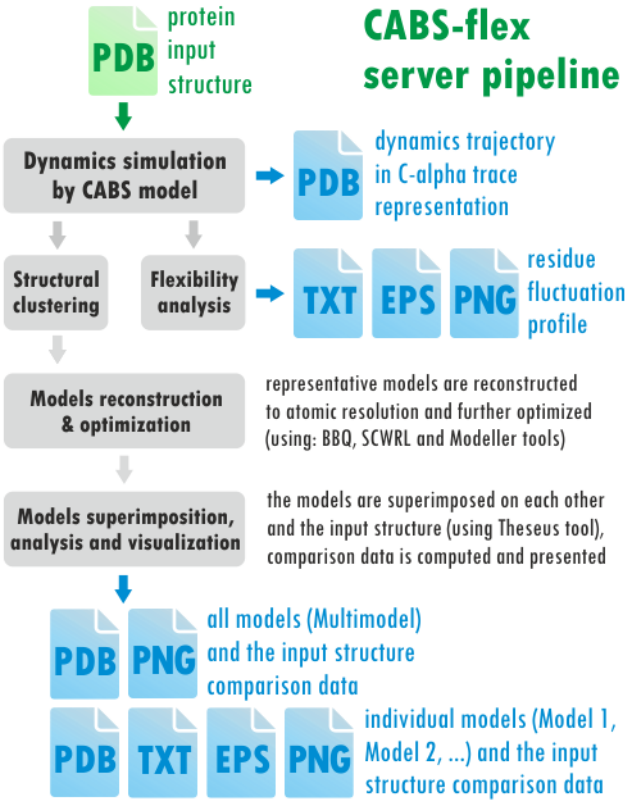
Wyniki przedstawione w pracy sugerują, że pole siłowe CABS — z rozpuszczalnikiem w formie niejawnej — pozwala symulować zachowanie łańcucha białkowego podobnie jak przy zastosowaniu dynamiki molekularnej.

Przesunięcie na wykresie pomiędzy funkcjami autokorelacji z trajektorii CABS a średnią z trajektorii MD pozwala wysnuć wniosek, że zastosowany gruboziarnisty model skraca czas pracy procesora około $6 \times 10^3$-krotnie[2]. Nachylenie krzywej świadczy o tym samym typie dyfuzji występującej po zastosowaniu poszczególnych pól siłowych.

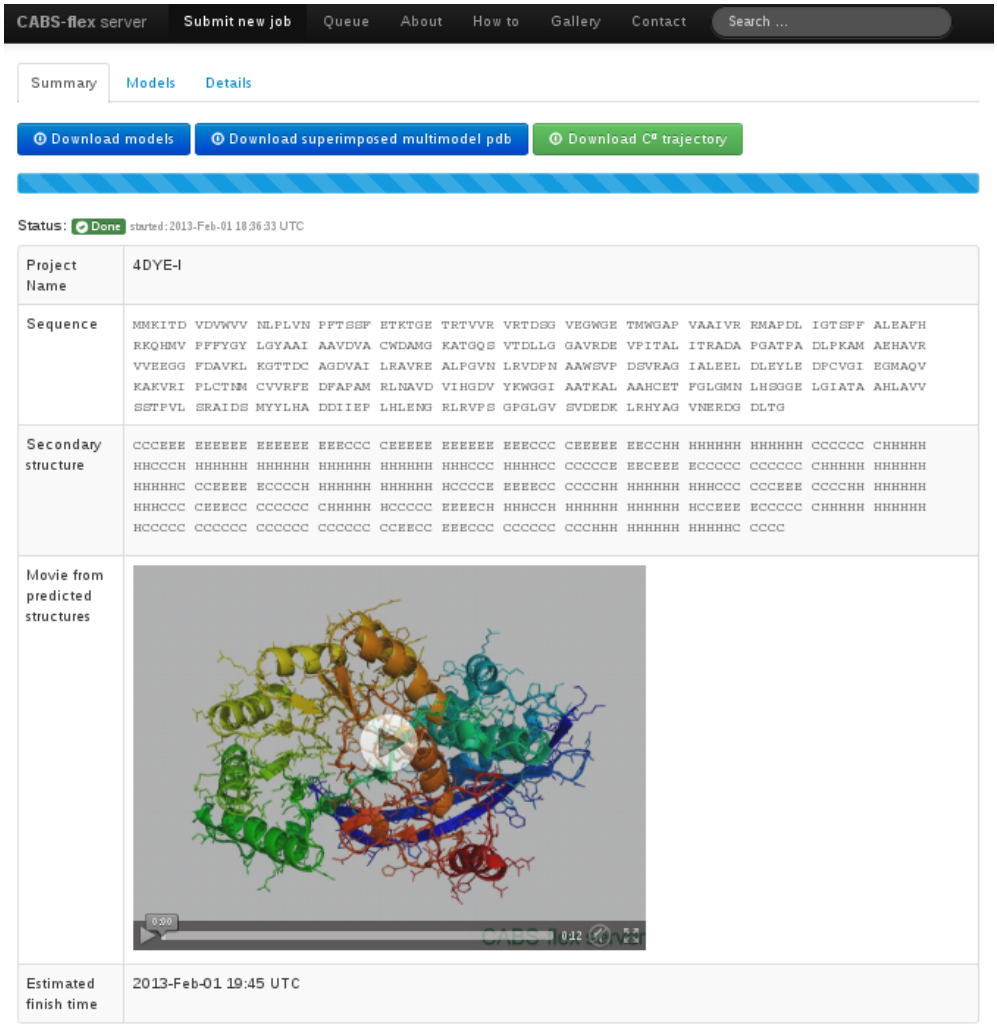
Kontynuując pracę w tym kierunku, przygotowałem aplikację umożliwiającą modelowanie dynamiki CABS dostępną z poziomu strony internetowej. Użytkownik, podając kod PDB wybranego białka (bądź przesyłając własny łańcuch w formacie PDB) ma możliwość przeprowadzenia symulacji na serwerach *Pracowni Teorii Biopolimerów* (metodyką opisaną w Pracy E), otrzymując zbiór konformacji alternatywnych w reprezentacji pełno-atomowej, trajektorię C$\alpha$ oraz inne dane, jak np. wartości fluktuacji atomów w funkcji indeksu reszty aminokwasowej, nałożenia konformacji alternatywnych na strukturę odniesienia, itp. Przepływ danych przedstawiłem na Rysunku 18.

2 Należy jednak pamiętać o tym, że ze względu na podejście stochastyczne, *czas biologiczny* modelu CABS będzie czasem przybliżonym, tj. liczba kroków MC przypadająca na jednostkę czasu będzie różna dla różnych białek.

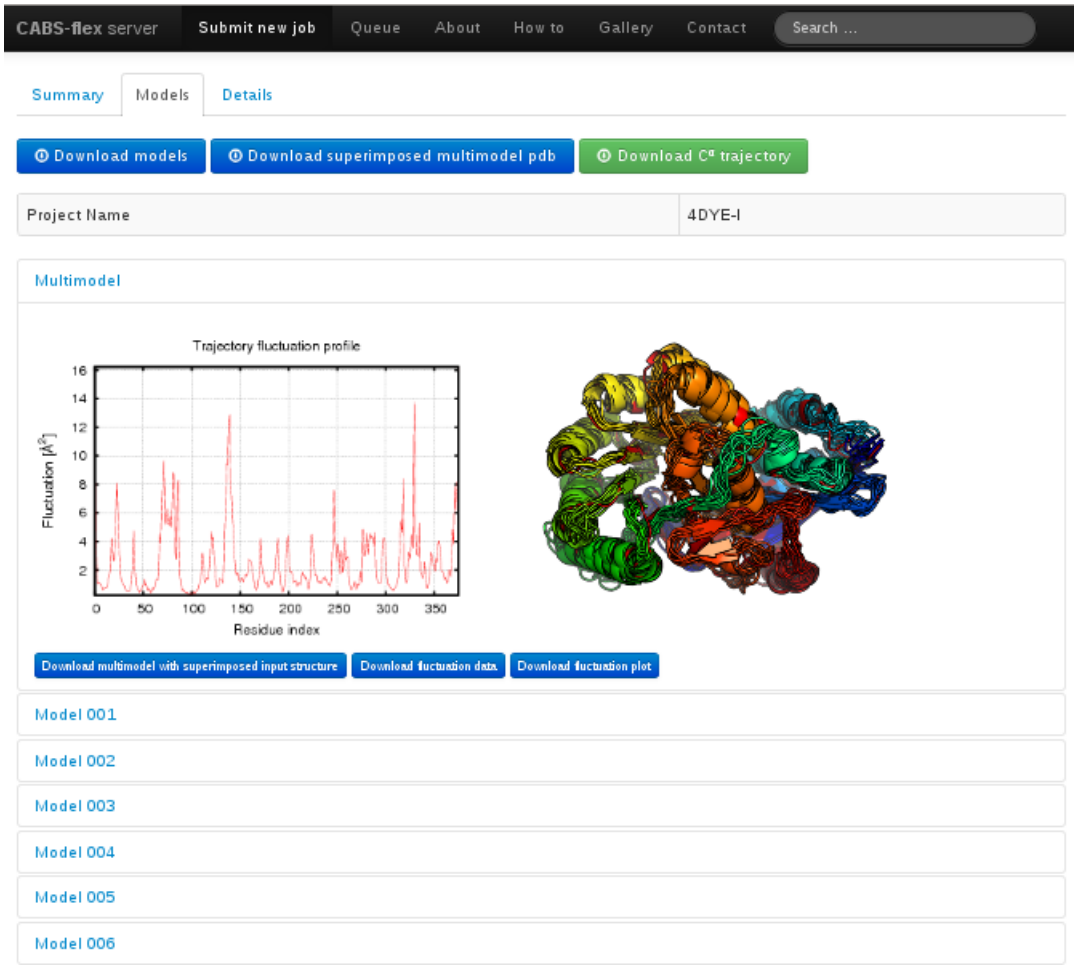
Aplikacja dostępna jest dla użytkowników pod adresem http://biocomp-.chem.uw.edu.pl/CABSflex. Przykładowy wynik pracy aplikacji przedstawiłem na Rysunkach 19–21.



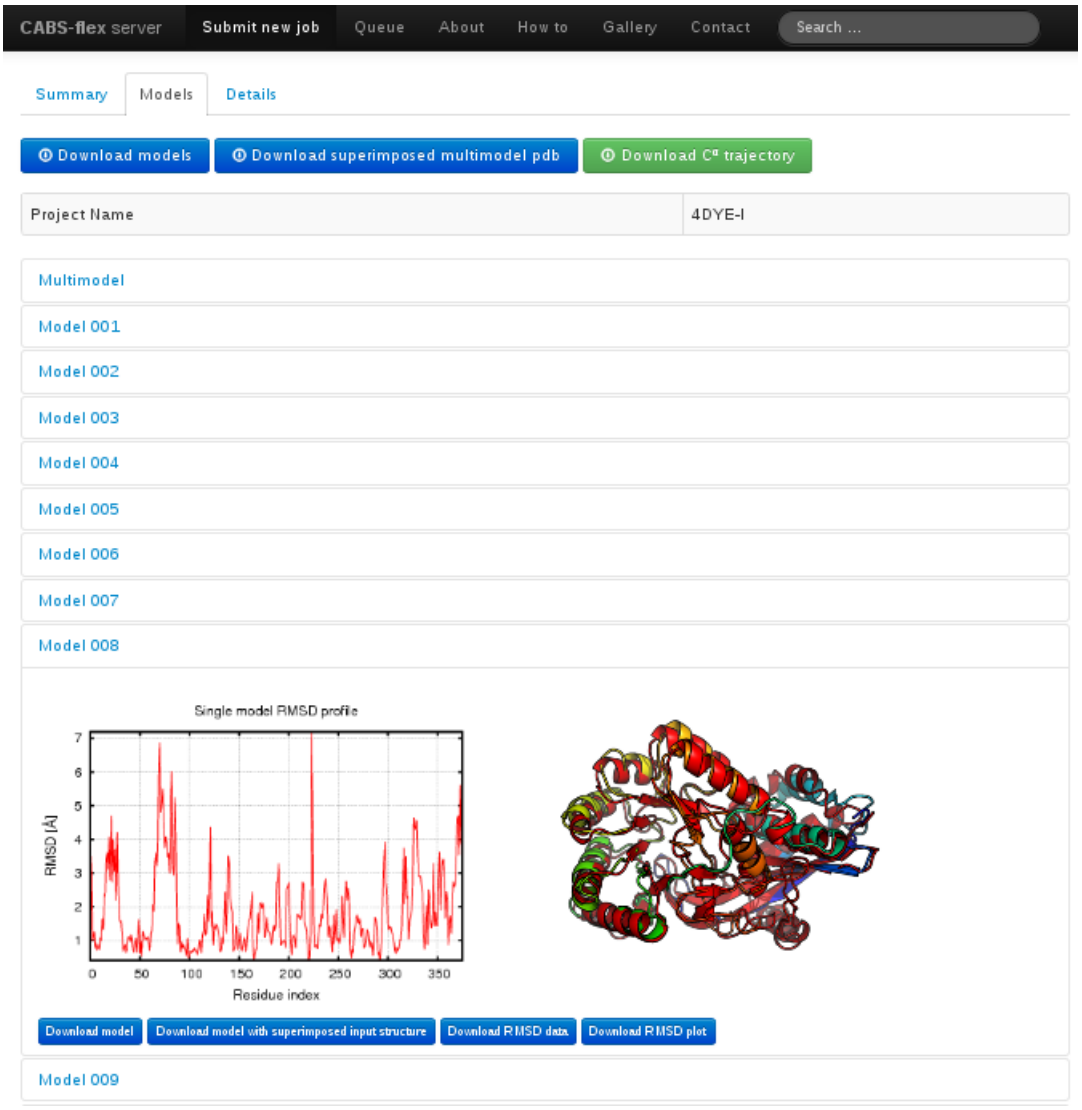

**Rysunek 18:** Przepływ danych i zastosowane kroki modelowania w opracowanej procedurze przewidywania wartości amplitudy fluktuacji atomów C$\alpha$ w strukturach białek z wykorzystaniem gruboziarnistego modelu CABS.




**Rysunek 19:** Strona wynikowa aplikacji CABS-*flex server*. Użytkownik, po wprowadzeniu kodu PDB białka i odczekaniu około godziny czasu trwania symulacji, otrzymuje informację o sekwencji, przypisanej strukturze drugorzędowej oraz pseudo-trajektorię atomów C$\alpha$ i zbiór kilkunastu reprezentatywnych modeli pełno-atomowych.




**Rysunek 20:** Strona wynikowa aplikacji CABS-*flex server*. Powyższy zrzut ekranu przedstawia nałożone na siebie konformacje alternatywne białka 4dye oraz wykres fluktuacji atomów C$\alpha$.




**Rysunek 21:** Strona wynikowa aplikacji CABS-*flex server*. Wykres przedstawia RMSD dla poszczególnych reszt po nałożeniu otrzymanego alternatywnego modelu na model otrzymany doświadczalnie.

# Część IV

# Podsumowanie

# 9 WNIOSKI KOŃCOWE

W niniejszej rozprawie pokazałem zastosowanie metod gruboziarnistych w modelowaniu struktury (CABS) i dynamiki (CABS, GNM) białek.

W trakcie badań powstały zoptymalizowane narzędzia ułatwiające modelowanie porównawcze struktury białek (na podstawie sekwencji białka celu i struktur-szablonów białek podobnych sekwencyjnie) oraz modelowanie *de novo* (na podstawie samej sekwencji-celu). Wykazałem ponadto, że model CABS nadaje się do modelowania struktur fragmentów białkowych (pętli) równie dobrze lub lepiej niż inne popularne algorytmy, wykorzystujące mniej zredukowane modele struktury (ROSETTA, MODELLER).

Zbadałem również zależność dynamiki białka od jego struktury i opracowałem unikalny model służący do przewidywania wartości bezwzględnych amplitud fluktuacji atomów C$\alpha$ w oparciu o strukturę.

Kierując się pragmatyzmem, starałem się udostępniać opracowane narzędzia społeczności naukowej, przez co większość metod jest ogólnodostępna w internecie w formie łatwych w użyciu interfejsów opartych na serwerach HTTP bądź jako kody programów na licencjach wolnego oprogramowania.

# 10 BIBLIOGRAFIA


Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., i Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

Anderson, D. P. (2004). BOINC: A System for Public-Resource Computing and Storage. In *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, GRID '04, pages 4–10, Washington, DC, USA. IEEE Computer Society.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(96):223–30.

Bahar, I., Atilgan, A. R., i Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & design*, 2(3):173–81.

Bahar, I. i Rader, a. J. (2005). Coarse-grained normal mode analysis in structural biology. *Current opinion in structural biology*, 15(5):586–92.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., i Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, 28(1):235–42.

Blaszczyk, M., Jamroz, M., Gront, D., i Kolinski, A. (2012). Protein Structure Prediction Using CABS – A Consensus Approach. In Carloni, P., Hansmann, U., Lippert, T., Meinke, J., Mohanty, S., Nadler, W., i Zimmermann, O., editors, *From Computational Biophysics to Systems Biology (CBSB11) Proceedings, IAS Series: Vol. 8*, pages 29–32, Jülich.

Boehr, D. D., Dyson, H. J., i Wright, P. E. (2006). An NMR perspective on enzyme dynamics. *Chemical reviews*, 106(8):3055–79.

Bornot, A., Etchebest, C., i de Brevern, A. G. (2011). Predicting protein flexibility through the prediction of local structures. *Proteins*, 79(3):839–52.

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L.,

Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., i Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–614.

Brüschweiler, R. (2003). Efficient RMSD measures for the comparison of two molecular ensembles. Root-mean-square deviation. *Proteins*, 50(1):26–34.

Brüschweiler, R. i Wright, P. E. (1994). NMR Order Parameters of Biomolecules: A New Analytical Representation and Application to the Gaussian Axial Fluctuation Model. *Journal of the American Chemical Society*, 116(18):8426–8427.

Buch, I., Giorgino, T., i De Fabritiis, G. (2011). Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25):10184–9.

Bujnicki, J. M., Elofsson, A., Fischer, D., i Rychlewski, L. (2001). Structure prediction meta server. *Bioinformatics (Oxford, England)*, 17(8):750–1.

Canutescu, A. A. i Dunbrack, R. L. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein science : a publication of the Protein Society*, 12(5):963–72.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., i Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–88.

Chakravarty, S. i Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*, 7(7):723–32.

Chiappori, F., Merelli, I., Colombo, G., Milanesi, L., i Morra, G. (2012). Molecular Mechanism of Allosteric Communication in Hsp70 Revealed by Molecular Dynamics Simulations. *PLoS Computational Biology*, 8(12):e1002844.

Chiti, F. i Dobson, C. M. (2009). Amyloid formation by globular proteins under native conditions. *Nature chemical biology*, 5(1):15–22.

Chow, E., Rendleman, C. A., Bowers, K. J., Dror, R. O., Hughes, D. H., Gullingsrud, J., Sacerdoti, F. D., i Shaw, D. E. (2008a). Anton, a special-purpose

machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91.

Chow, E., Rendleman, C. A., Bowers, K. J., Dror, R. O., Hughes, D. H., Gullingsrud, J., Sacerdoti, F. D., i Shaw, D. E. (2008b). Desmond performance on a cluster of multicore processors. Technical report, D. E. Shaw Research Technical Report DESRES/TR–2008-01.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., i Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117(19):5179–5197.

Cortes, C. i Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Damm, K. L. i Carlson, H. A. (2006). Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophysical journal*, 90(12):4558–73.

Day, R. N. i Schaufele, F. (2008). Fluorescent protein tools for studying protein dynamics in living cells: a review. *Journal of biomedical optics*, 13(3):031202.

De Simone, A., Montalvao, R. W., i Vendruscolo, M. (2011). Determination of Conformational Equilibria in Proteins Using Residual Dipolar Couplings. *Journal of chemical theory and computation*, 7(12):4189–4195.

Debe, D. A., Danzer, J. F., Goddard, W. A., i Poleksic, A. (2006). STRUCTFAST: protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. *Proteins*, 64(4):960–7.

Doniach, S. (2001). Changes in biomolecular conformation seen by small angle X-ray scattering. *Chemical reviews*, 101(6):1763–78.

Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., i Shaw, D. E. (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41:429–52.

Eisenmesser, E. Z., Bosco, D. A., Akke, M., i Kern, D. (2002). Enzyme dynamics during catalysis. *Science*, 295(5559):1520–3.

Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E., i Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–21.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U., i Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]*, Chapter 2:Unit 2.9.

Feng, Y., Kloczkowski, A., i Jernigan, R. L. (2010). Potentials 'R' Us web-server for protein energy estimations with coarse-grained knowledge-based potentials. *BMC bioinformatics*, 11:92.

Fiaux, J., Bertelsen, E. B., Horwich, A. L., i Wüthrich, K. (2002). NMR analysis of a 900K GroEL GroES complex. *Nature*, 418(6894):207–11.

Finkelstein, A. V., Badretdinov AYa, i Gutin, A. M. (1995). Why do protein architectures have Boltzmann-like statistics? *Proteins*, 23(2):142–50.

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993.

Fitter, J., Katranidis, A., Rosenkranz, T., Atta, D., Schlesinger, R., i Büldt, G. (2011). Single molecule fluorescence spectroscopy: a tool for protein studies approaching cellular environmental conditions. *Soft Matter*, 7(4):1254.

Frederick, K. K., Marlow, M. S., Valentine, K. G., i Wand, A. J. (2007). Conformational entropy in molecular recognition by proteins. *Nature*, 448(7151):325–9.

Gopal, S. M., Mukherjee, S., Cheng, Y.-M., i Feig, M. (2010). PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins*, 78(5):1266–81.

Górecki, A., Szypowski, M., Długosz, M., i Trylska, J. (2009). RedMD–reduced molecular dynamics package. *Journal of computational chemistry*, 30(14):2364–73.

Gu, J., Gribskov, M., i Bourne, P. E. (2006). Wiggle-predicting functionally flexible regions from primary sequence. *PLoS computational biology*, 2(7):e90.

Gunasekaran, K., Ma, B., i Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 57(3):433–43.

Guvench, O. i MacKerell, A. D. (2008). Comparison of protein force fields for molecular dynamics simulations. *Methods in molecular biology (Clifton, N.J.)*, 443:63–88.

Hajdu, J., Neutze, R., Sjögren, T., Edman, K., Szöke, A., Wilmouth, R. C., i Wilmot, C. M. (2000). Analyzing protein functions in four dimensions.

*Nature structural biology*, 7(11):1006–12.

Haliloglu, T. i Bahar, I. (1999). Structure-based analysis of protein dynamics: Comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. *Proteins: Structure, Function, and Genetics*, 37(4):654–667.

Haliloglu, T., Bahar, I., i Erman, B. (1997). Gaussian Dynamics of Folded Proteins. *Physical Review Letters*, 79(16):3090–3093.

Halle, B. (2002). Flexibility and packing in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3):1274–9.

Hamelryck, T. (2005). An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins*, 59(1):38–48.

Hammes, G. G., Benkovic, S. J., i Hammes-Schiffer, S. (2011). Flexibility, diversity, and cooperativity: pillars of enzyme catalysis. *Biochemistry*, 50(48):10422–30.

Harvey, M. J., Giupponi, G., i Fabritiis, G. D. (2009). ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *Journal of Chemical Theory and Computation*, 5(6):1632–1639.

Hess, B., Kutzner, C., van der Spoel, D., i Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447.

Hirose, S., Yokota, K., Kuroda, Y., Wako, H., Endo, S., Kanai, S., i Noguchi, T. (2010). Prediction of protein motions from amino acid sequence and its application to protein-protein interaction. *BMC structural biology*, 10:20.

Homans, S. W. (2004). NMR spectroscopy tools for structure-aided drug design. *Angewandte Chemie (International ed. in English)*, 43(3):290–300.

Homans, S. W. (2005). Probing the binding entropy of ligand-protein interactions by NMR. *Chembiochem : a European journal of chemical biology*, 6(9):1585–91.

Huang, Y. J. i Montelione, G. T. (2005). Structural biology: proteins flex to function. *Nature*, 438(7064):36–7.

Hubbard, T. J. (1999). RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. *Proteins*, Suppl 3(S3):15–21.

Jamroz, M. (2008). *Opracowanie algorytmu do odbudowy grup bocznych w uproszczonych modelach białek*. Praca magisterska, Uniwersytet Warszawski.

Jamroz, M. i Kolinski, A. (2010). Modeling of loops in proteins: a multi-method approach. *BMC structural biology*, 10(1):5.

Jamroz, M., Kolinski, A., i Kihara, D. (2012). Structural features that predict real-value fluctuations of globular proteins. *Proteins*, 80(5):1425–35.

Jamroz, M., Orozco, M., Kolinski, A., i Kmiecik, S. (2013). A Consistent View of Protein Fluctuations from All-atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-based Force-field. *Journal of Chemical Theory and Computation*, 9(1):119–125.

Johnston, J. M. i Filizola, M. (2011). Showcasing modern molecular dynamics simulations of membrane proteins through G protein-coupled receptors. *Current opinion in structural biology*, 21(4):552–8.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202.

Jorgensen, W. L. i Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923.

Kabsch, W. i Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637.

Kaczanowski, S. i Zielenkiewicz, P. (2009). Why similar protein sequences encode similar three-dimensional structures? *Theoretical Chemistry Accounts*, 125(3-6):643–650.

Kaur, J., Bhardwaj, A., Huang, Z., i Knaus, E. E. (2012). Aspirin analogues as dual cyclooxygenase-2/5-lipoxygenase inhibitors: synthesis, nitric oxide release, molecular modeling, and biological evaluation as anti-inflammatory agents. *ChemMedChem*, 7(1):144–50.

Kern, D. i Zuiderweg, E. R. P. (2003). The role of dynamics in allosteric regulation. *Current opinion in structural biology*, 13(6):748–57.

Kihara, D., Zhang, Y., Lu, H., Kolinski, A., i Skolnick, J. (2002). Ab initio protein structure prediction on a genomic scale: application to the Mycoplasma genitalium genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5993–8.

Kmiecik, S., Jamróz, M., Zwolińska, A., Gniewek, P., i Kolinski, A. (2008). Designing an Automatic Pipeline for Protein Structure Prediction. In Hansmann, U., Meinke, J., Mohanty, S., Nadler, W., i Zimmermann, O., editors, *From Computational Biophysics to Systems Biology (CBSB08) Proceedings, NIC Series Vol. 40*, volume 40, pages 105–108, Jülich.

Kmiecik, S. i Kolinski, A. (2007). Characterization of protein-folding pathways by reduced-space modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):12330–5.

Kolinski, A. (2004a). Protein modeling and structure prediction with a reduced representation. *Acta biochimica Polonica*, 51(2):349–71.

Kolinski, A. (2004b). Protein modeling and structure prediction with a reduced representation. *Acta biochimica Polonica*, 51(2):349–71.

Kolinski, A. i Skolnick, J. (1998). Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. *Proteins*, 32(4):475–94.

Kolinski, A. i Skolnick, J. (2004). Reduced models of proteins and their applications. *Polymer*, 45(2):511–524.

Koliński, A. i Bujnicki, J. M. (2005). Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, 61 Suppl 7:84–90.

Kondrashov, D. a., Van Wynsberghe, A. W., Bannen, R. M., Cui, Q., i Phillips, G. N. (2007). Protein structural variation in computational models and crystallographic data. *Structure*, 15(2):169–77.

Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98–104.

Kryshtafovych, A., Fidelis, K., i Moult, J. (2011). CASP9 results compared to those of previous CASP experiments. *Proteins*, 79 Suppl 1:196–207.

Kundu, S., Melton, J. S., Sorensen, D. C., i Phillips, G. N. (2002). Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophysical journal*, 83(2):723–32.

Kurowski, M. i Bujnicki, J. (2003). GeneSilico protein structure prediction meta-server. *Nucleic Acids Research*, 31(13):3305–3307.

Kyte, J. i Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–32.

Lane, T. J., Shukla, D., Beauchamp, K. a., i Pande, V. S. (2012). To milliseconds and beyond: challenges in the simulation of protein folding. *Current Opinion in Structural Biology*, pages 1–8.

Lasker, K., Förster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A., i Baumeister, W. (2012). Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proceedings of the National Academy of Sciences of the United States of America*, 109(5):1380–7.

Lassila, J. K. (2010). Conformational diversity and computational enzyme design. *Current opinion in chemical biology*, 14(5):676–82.

Lee, C. i Ham, S. (2011). Characterizing amyloid-beta protein misfolding from molecular dynamics simulations with explicit water. *Journal of computational chemistry*, 32(2):349–55.

Lee, S., Chen, M., Yang, W., i Richards, N. G. J. (2010). Sampling long time scale protein motions: OSRW simulation of active site loop conformational free energies in formyl-CoA:oxalate CoA transferase. *Journal of the American Chemical Society*, 132(21):7252–3.

Li, S. C., Bu, D., Xu, J., i Li, M. (2011). Finding nearly optimal GDT scores. *Journal of computational biology a journal of computational molecular cell biology*, 18(5):693–704.

Li, W., Zhang, Y., Kihara, D., Huang, Y. J., Zheng, D., Montelione, G. T., Kolinski, A., i Skolnick, J. (2003). TOUCHSTONEX: protein structure prediction with sparse NMR data. *Proteins*, 53(2):290–306.

Lill, M. A. (2011). Efficient incorporation of protein flexibility and dynamics into molecular docking simulations. *Biochemistry*, 50(28):6157–69.

Lin, C.-P., Huang, S.-W., Lai, Y.-L., Yen, S.-C., Shih, C.-H., Lu, C.-H., Huang, C.-C., i Hwang, J.-K. (2008). Deriving protein dynamical properties from weighted protein contact number. *Proteins*, 72(3):929–35.

Lindorff-Larsen, K., Best, R. B., Depristo, M. a., Dobson, C. M., i Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics.

*Nature*, 433(7022):128–32.

Lindorff-Larsen, K., Piana, S., Dror, R. O., i Shaw, D. E. (2011). How Fast-Folding Proteins Fold. *Science*, 334(6055):517–520.

Liwo, A., Khalili, M., i Scheraga, H. A. (2005). Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2362–7.

Lu, W. C., Wang, C. Z., Yu, E. W., i Ho, K. M. (2006). Dynamics of the trimeric AcrB transporter protein inferred from a B-factor analysis of the crystal structure. *Proteins*, 62(1):152–8.

Lundström, J., Rychlewski, L., Bujnicki, J., i Elofsson, A. (2001). Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein science : a publication of the Protein Society*, 10(11):2354–62.

Ma, B., Kumar, S., Tsai, C. J., i Nussinov, R. (1999). Folding funnels and binding mechanisms. *Protein engineering*, 12(9):713–20.

Mandell, D. J. i Kortemme, T. (2009). Backbone flexibility in computational protein design. *Current opinion in biotechnology*, 20(4):420–8.

Meinhold, L. i Smith, J. C. (2005). Fluctuations and correlations in crystalline protein dynamics: a simulation analysis of staphylococcal nuclease. *Biophysical journal*, 88(4):2554–63.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., i Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087.

Mittermaier, A. i Kay, L. E. (2006). New tools provide new insights in NMR studies of protein dynamics. *Science (New York, N.Y.)*, 312(5771):224–8.

Moffat, K. (1998). Time-Resolved Crystallography. *Acta Crystallographica Section A Foundations of Crystallography*, 54(6):833–841.

Monod, J., Wyman, J., i Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *Journal of molecular biology*, 12:88–118.

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., i Marrink, S.-J. (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834.

Moritsugu, K. i Smith, J. C. (2007). Coarse-grained biomolecular simulation with REACH: realistic extension algorithm via covariance Hessian. *Biophy-*

*sical journal*, 93(10):3460–9.

Mozziconacci, J.-C., Arnoult, E., Bernard, P., Do, Q. T., Marot, C., i Morin-Allory, L. (2005). Optimization and validation of a docking-scoring protocol; application to virtual screening for COX-2 inhibitors. *Journal of medicinal chemistry*, 48(4):1055–68.

Orengo, C. a., Michie, a. D., Jones, S., Jones, D. T., Swindells, M. B., i Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure (London, England : 1993)*, 5(8):1093–108.

Palmer, A. G., Kroenke, C. D., i Loria, J. P. (2001). Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods in enzymology*, 339:204–38.

Pan, X.-Y. i Shen, H.-B. (2009). Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. *Protein and peptide letters*, 16(12):1447–54.

Phillips, G. N. (1990). Comparison of the dynamics of myoglobin in different crystal forms. *Biophysical journal*, 57(2):381–3.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., i Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 26(16):1781–802.

Piela, L. (2005). *Idee chemii kwantowej*. Wydawnictwo Naukowe PWN, Warszawa.

Ponder, J. W. i Richards, F. M. (1987). An efficient newton-like method for molecular mechanics energy minimization of large molecules. *Journal of Computational Chemistry*, 8(7):1016–1024.

Rasmussen, B. F., Stock, A. M., Ringe, D., i Petsko, G. A. (1992). Crystalline ribonuclease A loses function below the dynamical transition at 220 K. *Nature*, 357(6377):423–4.

RCSB Protein DataBank (2012). RCSB PDB - Content Growth Report.

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., i Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in enzymology*, 383:66–93.

Rossi, K. a., Weigelt, C. a., Nayeem, A., i Krystek, S. R. (2007). Loopholes and missing links in protein modeling. *Protein science : a publication of the Protein Society*, 16(9):1999–2012.

Rueda, M., Ferrer-Costa, C., Meyer, T., Perez, A., Camps, J., Hospital, A., Gelpi, J. L., Orozco, M., Pérez, A., i Gelpí, J. L. (2007). A consensus view of protein dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 104(3):796–801.

Sali, A. i Blundell, T. (1993). Comparative protein modeling by satisfaction of spatial restraints. *Journal of molecular biology*, 234:779–815.

Salmon, L., Bouvignies, G., Markwick, P., i Blackledge, M. (2011). Nuclear magnetic resonance provides a quantitative description of protein conformational flexibility on physiologically important time scales. *Biochemistry*, 50(14):2735–47.

Sanejouand, Y.-H. (2013). Elastic Network Models: Theoretical and Empirical Foundations. In Monticelli, L. i Salonen, E., editors, *Biomolecular Simulations*, volume 924 of *Methods in Molecular Biology*, pages 601–616. Humana Press.

Sanner, M. F., Olson, A. J., i Spehner, J. C. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–20.

Schlessinger, A. i Rost, B. (2005). Protein flexibility and rigidity predicted from sequence. *Proteins*, 61(1):115–26.

Schlessinger, A., Yachdav, G., i Rost, B. (2006). PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, 22(7):891–3.

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., i Wriggers, W. (2010). Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science*, 330(6002):341–346.

Shen, M.-Y. i Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein science : a publication of the Protein Society*, 15(11):2507–2524.

Shih, C.-h., Huang, S.-W., Yen, S.-C., Lai, Y.-L., Yu, S.-H., i Hwang, J.-K. (2007). A simple way to compute protein dynamics without a mechanical model. *Proteins*, 68(1):34–8.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Current opinion in structural biology*, 5(2):229–35.

Skolnick, J., Zhang, Y., Arakaki, A. K., Kolinski, A., Boniecki, M., Szilágyi, A., i Kihara, D. (2003). TOUCHSTONE: a unified approach to protein structure prediction. *Proteins*, 53 Suppl 6:469–79.

Smith, D. K., Radivojac, P., Obradovic, Z., Dunker, A. K., i Zhu, G. (2003). Improved amino acid flexibility parameters. *Protein science : a publication of the Protein Society*, 12(5):1060–72.

Sugita, Y. i Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151.

Sułkowska, J. I. (2007). *Zwijanie i rozwijanie białek w modelach gruboziarnistych*. Praca doktorska, Instytut Fizyki PAN.

Swendsen, R. i Wang, J. (1986). Replica Monte Carlo simulation of spin glasses. *Physical review letters*, 57(21):2607–2609.

Taiji, M., Narumi, T., Ohno, Y., Futatsugi, N., Suenaga, A., Takada, N., i Konagaya, A. (2003). Protein Explorer: A Petaflops Special-Purpose Computer System for Molecular Dynamics Simulations. In *Proceedings of the 2003 ACM/IEEE conference on Supercomputing*, SC '03, page 15, New York, NY, USA. ACM.

Taketomi, H., Ueda, Y., i Go, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. *International Journal of Peptide and Protein Research*, 7(6):445–459.

Teilum, K., Olsen, J. G., i Kragelund, B. B. (2009). Functional aspects of protein flexibility. *Cellular and molecular life sciences : CMLS*, 66(14):2231–47.

Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical review letters*, 77(9):1905–1908.

Trylska, J. (2010). Coarse-grained models to study dynamics of nanoscale biomolecules and their applications to the ribosome. *Journal of physics. Condensed matter : an Institute of Physics journal*, 22(45):453101.

Vane, J. R. (1971). Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs. *Nature: New biology*, 231(25):232–5.

Vendruscolo, M. i Dobson, C. M. (2011). Protein dynamics: Moore's law in molecular biology. *Current biology : CB*, 21(2):R68–70.

Vugmeyster, L. i Ostrovsky, D. (2011). Temperature dependence of fast carbonyl backbone dynamics in chicken villin headpiece subdomain. *Journal of biomolecular NMR*, 50(2):119–27.

Wallner, B. i Elofsson, A. (2005). Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics*, 21(23):4248–54.

Wang, C., Bradley, P., i Baker, D. (2007). Protein-protein docking with backbone flexibility. *Journal of molecular biology*, 373(2):503–19.

Wang, T., Frederick, K. K., Igumenova, T. I., Wand, A. J., i Zuiderweg, E. R. P. (2005). Changes in calmodulin main-chain dynamics upon ligand binding revealed by cross-correlated NMR relaxation measurements. *Journal of the American Chemical Society*, 127(3):828–9.

Wu, S., Skolnick, J., i Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC biology*, 5:17.

Yang, L., Song, G., i Jernigan, R. L. (2009). Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences of the United States of America*, 106(30):12347–52.

Yuan, Z., Zhao, J., i Wang, Z.-X. (2003). Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Engineering Design and Selection*, 16(2):109–114.

Zacharias, M. (2010). Accounting for conformational changes during protein-protein docking. *Current opinion in structural biology*, 20(2):180–6.

Zemla, A., Venclovas, C., Moult, J., i Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins*, Suppl 3(June):22–9.

Zhang, H., Zhang, T., Chen, K., Shen, S., Ruan, J., i Kurgan, L. (2009). On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins*, 76(3):617–36.

Zhang, Y., Kolinski, A., i Skolnick, J. (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophysical journal*, 85(2):1145–64.

# Prace stanowiące podstawę rozprawy

# A MODELING OF LOOPS IN PROTEINS: A MULTI-METHOD APPROACH

*Jamroz M, Kolinski A. (2010)* Modeling of loops in proteins: a multi-method approach. *BMC Structural Biology 10(1):5.*

Jamroz and Kolinski *BMC Structural Biology* 2010, **10**:5
http://www.biomedcentral.com/1472-6807/10/5

BMC Structural Biology

RESEARCH ARTICLE Open Access

# Modeling of loops in proteins: a multi-method approach

Michal Jamroz, Andrzej Kolinski*

## Abstract

**Background:** Template-target sequence alignment and loop modeling are key components of protein comparative modeling. Short loops can be predicted with high accuracy using structural fragments from other, not necessarily homologous proteins, or by various minimization methods. For longer loops multiscale approaches employing coarse-grained *de novo* modeling techniques should be more effective.

**Results:** For a representative set of protein structures of various structural classes test predictions of loop regions have been performed using MODELLER, ROSETTA, and a CABS coarse-grained *de novo* modeling tool. Loops of various length, from 4 to 25 residues, were modeled assuming an ideal target-template alignment of the remaining portions of the protein. It has been shown that classical modeling with MODELLER is usually better for short loops, while coarse-grained *de novo* modeling is more effective for longer loops. Even very long missing fragments in protein structures could be effectively modeled. Resolution of such models is usually on the level 2-6 Å, which could be sufficient for guiding protein engineering. Further improvement of modeling accuracy could be achieved by the combination of different methods. In particular, we used 10 top ranked models from sets of 500 models generated by MODELLER as multiple templates for CABS modeling. On average, the resulting molecular models were better than the models from individual methods.

**Conclusions:** Accuracy of protein modeling, as demonstrated for the problem of loop modeling, could be improved by the combinations of different modeling techniques.

## Background

Comparative modeling remains the most dependable and routinely used method for protein structure prediction [1,2]. The alternative term of homology modeling is frequently used. That is because the identification of a structural template (or templates) is typically based (although not always) on the homology relation between the target protein and the templates, which is usually reflected by a certain level of sequence similarity. When a template is being identified by some advanced Fold Recognition (FR) techniques, it is sometimes possible to identify templates that are structurally similar to the target without any obvious homology relations. This could be a genuine case of convergent evolution or (more frequently) the case when remote homology just can not be detected. Template free, *de novo* structure prediction is much more difficult and less dependable, although a steady progress is observed in this area of computational biology [3,4]. Most contemporary methods for *de novo* structure predictions heavily depend on certain aspects of evolutionary relationships between protein sequences and structures. The evolutionary methods are essential for the derivation of statistical potentials for *de novo* modeling and/or are employed in various strategies for extracting structure building blocks from known protein structures [5,6].

Classical homology modeling consists of three steps. First, a template for modeling needs to be identified and sequence alignment between the template and target sequences has to be generated. Usually, template identification is performed by certain standard tools, such as PSI-BLAST, and the resulting alignment is subsequently rectified by other tools and eventually by manual expert corrections. Remote templates can also be identified by FR procedures [7]. With the decreasing level of sequence similarity, which implies increasing evolutionary distance and thereby increasing structural differences

* Correspondence: kolinski@chem.uw.edu.pl
Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

Jamroz and Kolinski *BMC Structural Biology* 2010, **10**:5
http://www.biomedcentral.com/1472-6807/10/5


between the template and the target, alignments become more and more ambiguous. Accuracy of classical comparative modeling heavily relies on the fidelity of the template-target alignment.

In the second stage the aligned fragments of templates are used to generate the corresponding fragments of the target structure. In the simplest case of a single template only, this step reduces to mere copying the template coordinates according to the alignment. In the case of multiple templates a consensus scaffold could be built, for instance via the distribution of the spatial restrains read from the templates, as it is implemented in the MODELLER method [8]. The key component of this stage of modeling is construction of loop regions that are frequently missing in the template scaffold. In certain newer approaches to comparative modeling the entire structure of the target is built using templates as sources of restraints of various types [3,9]. The main aim and challenge of such approaches is to be able to build a model of the target structure which is more similar to the true structure of the target than to any of the templates used, especially for distant homology based modeling.

The third, and final, stage of modeling is structure refinement which involves repacking the side chains and energy minimization of the entire structure [10].

The above scheme, or its variants, of comparative modeling remains the best choice when significant fragments of the alignment are error-free, which is usually the case in the range of high level sequence similarity (e.g. 40%, or more, of identical residues in the alignment). In the "twilight zone" of low sequence similarity the alignments contain significant errors. These could be sometimes corrected by building a multi-template consensus modeling scaffold [11]. Alternatively, it is possible to design a completely different modeling schemes, in which the alignment is built simultaneously to the actual modeling process [12].

In this paper we address the issue of loop modeling, separating it from the alignment problem. The test set of proteins with missing loops consists of two sub-sets. The first subset, containing missing loops of 4-12 residues, has been taken from a recent work by Rossi, et al., excluding the cases of incomplete chains in the corresponding PDB entries [13]. The work provides a comprehensive comparison of loop modeling performance of the most popular comparative modeling software. The loop database employed in the work of Rossi was adapted from a compiled loop database assembled by Jacobson et al [13,14]. Additionally, the database used in this work was expanded with cases of much longer loops, up to 25 residues (the second sub-set). The database covers all the structural classes of proteins, with 186 internal loops of various length. The expanded range of the modeled loop lenghts addresses the possibility of the extension of the range of applicability and accuracy of challenging instances of comparative modeling. Four methods of loop modeling are compared in this work: MODELLER, ROSETTA, CABS and a combination of MODELLER with CABS. Since MODELLER is a commonly accepted reference standard in comparative modeling, the results are qualitatively (although indirectly) comparable with other approaches [13,15-20]. It should be noted that MODELLER is representative software for distance geometry and energy minimizations, while ROSETTA and CABS employ knowledge-based free search of a discretized conformational space. Thus, the comparison given in this paper should provide additional insights into the range of applicability of these qualitatively different approaches to protein molecular modeling. Previous computational experiments with the reconstruction of missing fragments of protein structures indicated that the coarse grained models (an early version of CABS and two other modeling tools based on similar principles) performed relatively well in the range of large fragments [21]. At this point we would like to present a comprehensive evaluation in a wide range of loop modeling instances.

## Results

For a representative test set of protein structures with missing loop fragments the loop reconstruction procedure was executed using MODELLER, ROSETTA, CABS and the MODELLER-CABS hybrid modeling pipeline. The test set is summarized in Table 1. Modeling procedures are described in the Methods section. The test proteins represent various structural classes, including mainly helical, beta and alpha/beta structures. All test structures are of high quality with resolution of at least 2 Å and the average temperature factor lower than 35. The missing loops are representative, and they are exposed to the solvent or partially buried, connecting various elements of the secondary structure. The modeled loops span a wide range of lengths, from 4 to 25 residues. This is a range that is relevant for standard comparative modeling. In several proteins more than one loop is modeled. In some cases the modeled loops can interact with one another, which can have some influence on the performance of respective methods.

Using MODELLER, we generated 500 examples of individual loop regions, which were subsequently ranked by the DOPE statistical potential [22]. Top ranked means the highest rank, while the "best" result means a structure that is closest to the actual experimental, structure of the loop. Similarly, ROSETTA models were ranked with ROSETTA potentials. CABS modeling provides a trajectory containing several hundred instances. These were subject to the clustering procedure.

Jamroz and Kolinski *BMC Structural Biology* 2010, **10**:5
http://www.biomedcentral.com/1472-6807/10/5



**Table 1 Protein codes and loop locations of test set of protein**

| Protein codes and loop locations of test set of protein. | |
|---|---|
| **loop length** | **PDB codes and loop ranges** |
| 4 | 7rsa 47-50, 4gcr 116-119, 2tgi 72-75, 2exo 161-164, 1xif 82-85, 1tml 42-45, 1tib 46-49, 1thw 194-197, 1rcf 111-114, 1ppn 42-45, 1plc 74-77, 1pbe 117-120, 1nfp 37-40, 1frd 59-62, 1cbs 21-24, 1ads 99-102, 1aaj 82-85 |
| 5 | 7rsa 75-79, 2hbg 37-41, 2cmd 188-192, 1vcc 63-67, 1tml 147-151, 1tca 157-161, 1sbp 181-185, 1prn 187-191, 1noa 88-92, 1nfp 95-99, 1nar 56-60, 1kuh 37-41, 1hbq 158-162, 1hbg 19-23, 1frd 83-87, 153l 131-135 |
| 6 | 5p21 104-109, 3pte 256-261, 3pte 131-136, 2ayh 81-86, 1tca 94-99, 1tca 38-43, 1rge 73-78, 1noa 25-30, 1mrp 233-238, 1gca 100-105, 1ede 180-185, 1cbs 66-71, 1brt 253-258, 1brt 174-179, 1ads 150-155, 1ads 149-154 |
| 7 | 5p21 83-89, 2pth 95-101, 1tml 20-26, 1tca 132-138, 1php 135-141, 1mbd 17-23, 1lif 64-70, 1iab 142-148, 1hbg 46-52, 1gca 196-202, 1edg 309-315, 1dad 116-122, 1brt 226-232, 1bkf 64-70, 1ads 186-192 |
| 8 | 2ayh 194-201, 1tml 187-194, 1thw 18-25, 1prn 150-157, 1nwp 84-91, 1nls 97-104, 1nar 192-199, 1hbq 31-38, 1arb 136-143, 1alc 34-41, 1ads 274-281 |
| 9 | 3pte 107-115, 2ayh 169-177, 1xnb 133-141, 1xnb 116-124, 1php 91-99, 1nls 131-139, 1ede 257-265, 1arb 168-176, 1aac 58-66 |
| 10 | 7rsa 87-96, 7rsa 33-42, 7rsa 110-119, 2cmd 57-66, 1whi 47-56, 1tca 23-32, 1scs 65-74, 1ppn 190-199, 1plc 42-51, 1mrj 173-182, 1ixh 84-93, 1gvp 49-58, 1fkf 63-72, 1arb 41-50, 1amp 181-190, 1ads 171-180, 1ads 170-179, 135l 18-27 |
| 11 | 3pte 91-101, 2pth 8-18, 1rcf 122-132, 1ixh 120-130, 1dad 42-52, 153l 154-164 |
| 12 | 2ayh 21-32, 1ixh 160-171, 1bkf 9-20, 1arb 74-85, 153l 98-109 |
| 16 | 1tml 73-88, 1tml 219-234, 1tca 184-199, 1rge 37-52, 1prn 106-121, 1nar 10-25, 1iab 136-151, 1frd 33-48, 1edg 233-248, 1edg 167-182, 1brt 57-72, 1amp 98-113, 1ads 210-225 |
| 18 | 1tml 73-90, 1tml 219-236, 1tca 184-201, 1prn 106-123, 1nar 10-27, 1iab 136-153, 1byt 807-824, 1byt 700-717, 1byt 359-376, 1byt 230-247, 1bst 57-74, 1bst 129-146, 1b57 209-226, 1awj 2-19, 1amp 98-115, 1ahj 101-118, 1ads 210-227, 1acc 36-53, 1acc 183-200 |
| 20 | 1br4 390-409, 1br4 349-368, 1br4 291-310, 1br2 246-265, 1azx 362-381 |
| 22 | 1tml 219-240, 1prn 106-127, 1nar 10-31, 1kk7 291-312, 1jez 117-138, 1itk 179-200, 1itk 157-178, 1e04 351-372, 1clq 380-401, 1br4 71-92, 1br4 256-277, 1b3k 322-343, 1aoa 182-203 |
| 23 | 1nfb 253-275, 1lzj 2-24, 1izl 21-43, 1i50 46-68, 1dzg 367-389 |
| 24 | 1uoz 224-247, 1mnd 277-300, 1miu 93-116, 1i19 415-438, 1hfb 86-109 |
| 25 | 2hs0 319-343, 2gah 437-461, 2fqf 293-317, 2e4y 311-335, 1zba 16-40, 1tml 219-243, 1qme 127-151, 1prn 106-130, 1kmh 117-141, 1eah 247-271, 1dms 596-620, 1dhx 376-400, 1dhx 11-35 |

Interestingly, in most cases the medoid from the entire simulation was closer to the true structure than the largest clusters' medoids. This suggest very good convergence of CABS simulations. Consequently, the medoid structures were reported as the top-ranked models.

The statistics of the results is shown in Figure 1, in which the loops of a given length are described by average values of cRMSD of the loop fragments (coordinate Root-Mean-Square Deviation) from the corresponding crystallographic structures. To extract the loop cRMSD values protein structures without the modeled loop fragments were superimposed and the deviation was computed only for the loops. In the entire text the values of cRMSD are reported for C$\alpha$ traces only. Corresponding data for all atom structures are essentially the same. The plots in Figure 1 clearly show two major trends. The first is obvious: with the increasing size of loops the average accuracy of modeling decreases. The second trend indicates, as expected, that the distribution of the quality of models, as measured by the difference between the best model and the top ranked models is much larger for MODELLER and ROSETTA as compared with CABS. For very long loops (20 and more residues) CABS results are on average better than for MODELLER and ROSETTA. The hybrid-CABS modeling takes advantages of different methods. Using top10 models generated by MODELLER the new method leads to results as good as MODELLER for short loops and noticeably better models in the range of long loops. When comparing with original CABS simulations the hybrid-CABS is much more accurate for short loops. This is illustrated in Figure 2. The results with hybrid-CABS show that there is always an added value in combining different modeling methods.

Figure 3 and Figure 4 show the distributions of cRMSD for 186 cases studied. The distribution is quite broad, especially for longer loops. Use of distinct modeling techniques increases chances for obtaining good quality models. Unfortunately, all methods produce results of scattered quality. The problem how to identify the cases for which the models are of good accuracy remains unsolved.

## Discussion

The loop modeling exercise described in this paper separates the two fundamental problems of comparative

Jamroz and Kolinski *BMC Structural Biology* 2010, **10**:5
http://www.biomedcentral.com/1472-6807/10/5



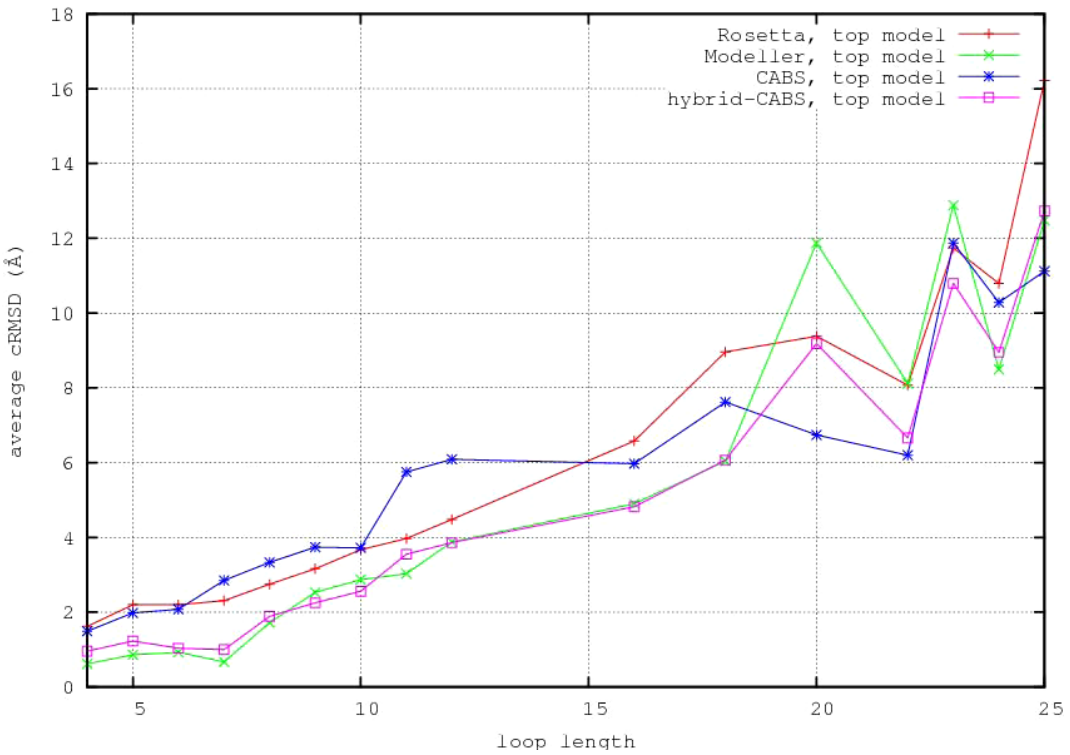



**Figure 1 Plots of average cRMSD of loops versus loop length**. Plots of average cRMSD of loops versus loop length. Best and top models generated by MODELLER, ROSETTA and CABS.

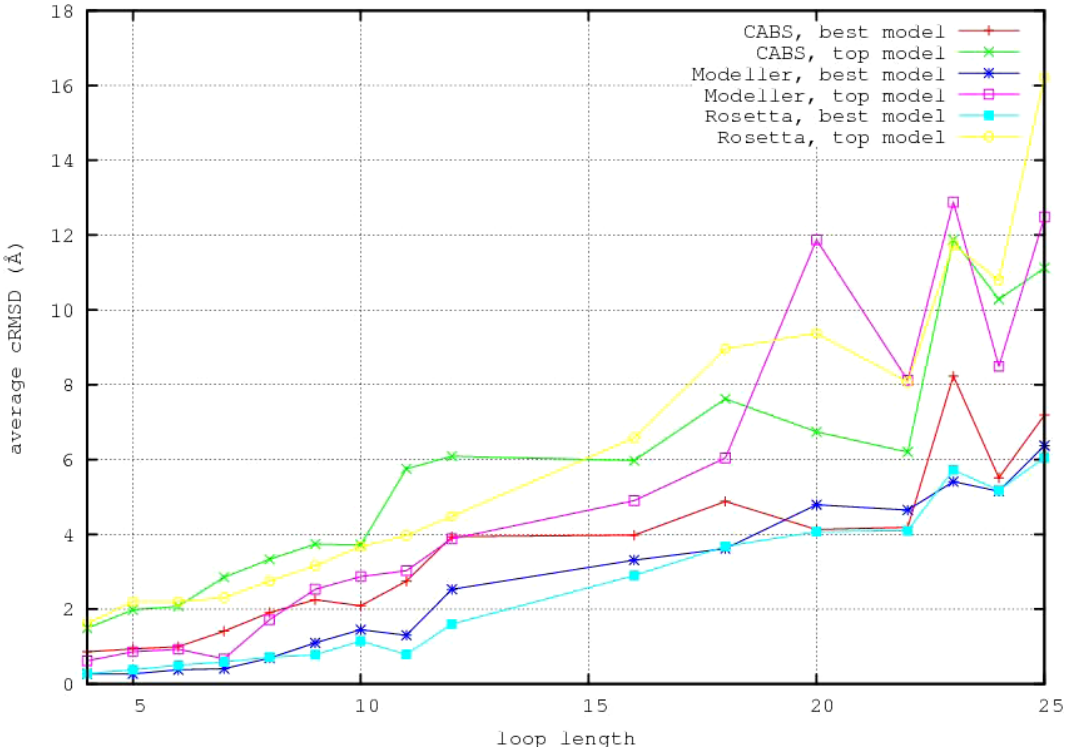



**Figure 2 Plots of average cRMSD of loops versus loop length**. Plots of average cRMSD of loops versus loop length. Top models generated by various modeling procedures, including MODELLER-CABS hybrid method (see the text).

Jamroz and Kolinski *BMC Structural Biology* 2010, **10**:5
http://www.biomedcentral.com/1472-6807/10/5



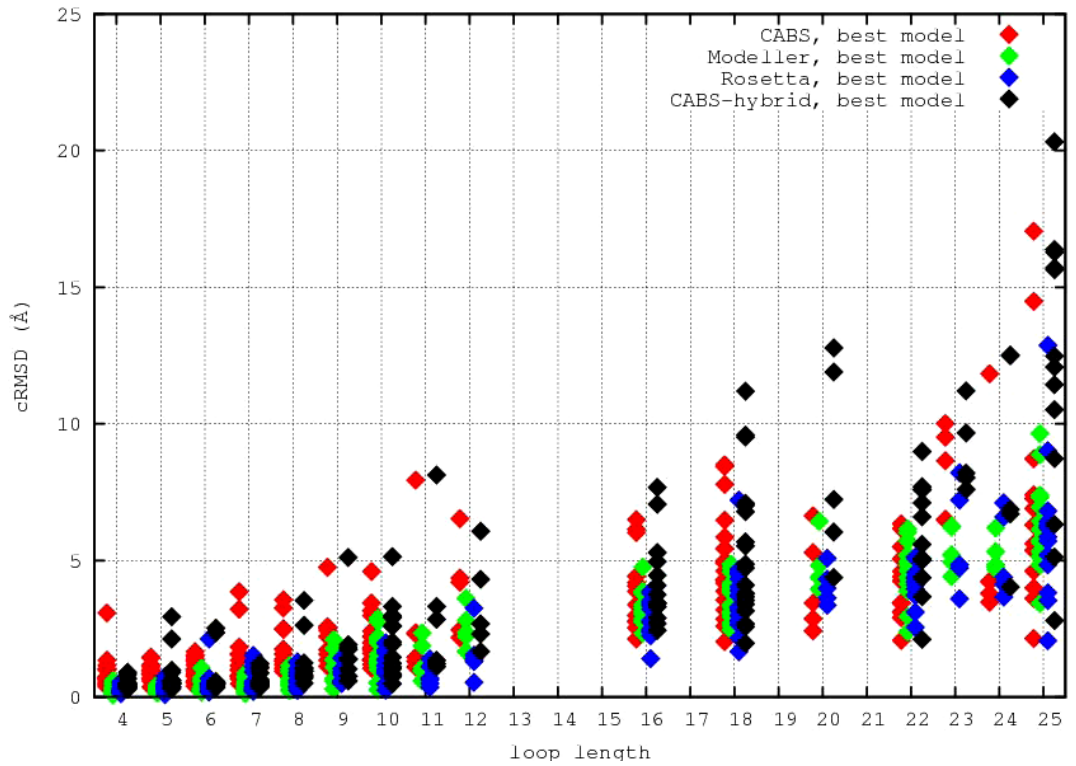



**Figure 3 Best loop models cRMSD**. Distribution of best loop models cRMSD for different modeling procedures.

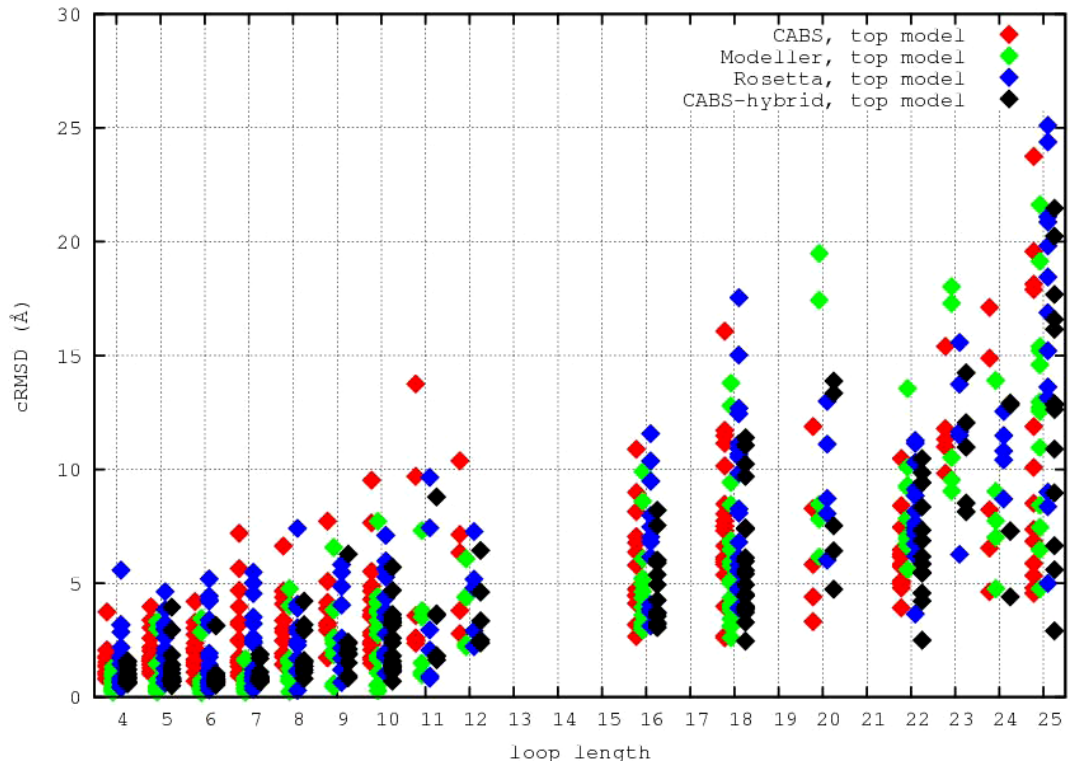



**Figure 4 Top ranked loop models cRMSD**. Distribution of top ranked loop models cRMSD for different modeling procedures.

Jamroz and Kolinski *BMC Structural Biology* 2010, **10**:5
http://www.biomedcentral.com/1472-6807/10/5


**Table 2 Summary of the average results from both modeling techniques.**

| | Average cRMSD (in Å) | | | |
|---|---|---|---|---|
| **Loop range** | **CABS top (best)** | **Modeller top (best)** | **Rosetta top (best)** | **Modeller-CABS hybrid top (best)** |
| 4-6 | **1.84 (0.93)** | **0.80 (0.31)** | **2.00 (0.38)** | 1.07 (0.66) |
| 7-12 | **3.83 (2.13)** | 2.20 **(1.10)** | **3.21 (0.89)** | 2.23 (1.75) |
| 16-25 | 8.11 **(5.23)** | 8.39 **(4.54)** | **10.02 (4.31)** | 7.87 (7.07) |

Average coordinate Root Mean-Square Deviation (cRMSD) from crystallographic structures in Ångstroms. Bold fonts indicates statistically relevant differences between individual methods and Modeller-CABS hybrid method (two-sample paired t-test, data in Additional file 1).

modeling: target-template alignment and modeling of missing fragments. Ideal alignments have been assumed and the excised loops reconstructed and compared with the native structures (cRMSD of the reconstructed loops after the superposition of the fixed parts of templates and models). As expected, MODELLER and ROSETTA proved to be more accurate for short loops, while CABS models were better for longer loops (see the compilation of cRMSD values for different ranges of loop sizes, shown in Table 2(two-sample paired t-test, data in Additional file 1)), although the difference is small. In spite of the coarse-grained character of the method, the models from CABS allow for the meaningful reconstruction of the side chain details for shorter, and therefore more accurately predicted, loops (see Figure 5). The predicted side-chains conformations, shown in Figure 5, are of crystallographic accuracy, except for the tail portion of one side-chain. For longer loops the side chains are less accurate and their native-like conformations and interaction patterns are observed only for the best models. Figure 6 shows a typical situation for the loops from the range of accuracy of 4-6 Å. In such cases the side chains are approximately at proper positions, although their conformations on the atomic level are not reproduced. Finally, it should be noted that the simulation results from CABS could be used for the analysis of loop dynamics. In recent publications we have shown that isothermal trajectories from CABS, executed at the folding transition temperature, reproduce folding mechanisms of small proteins very well [23,24]. Thus loop mobility could also be modeled. In order to obtain the best possible model of the lowest energy structures, in the present study we used Replica Exchange Monte Carlo. Thus, the dynamics of the system is artificial. Obviously, isothermal simulations could be performed for the models obtained, leading to meaningful description of loop mobility. This was, however, beyond the scope of the present work.

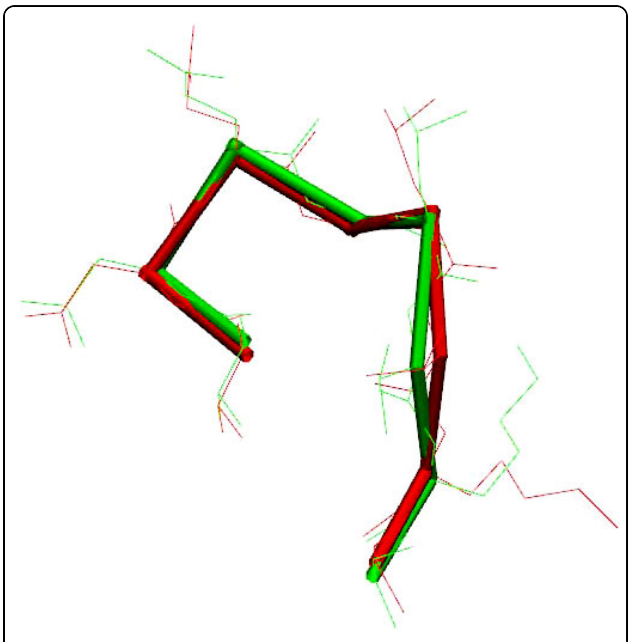


**Figure 5 Superimposition**. Sidechains of 149-154 loop from 1ads crystallographic structure (red) with superimposed CABS (cRMSD 0.70 Å) model (green).

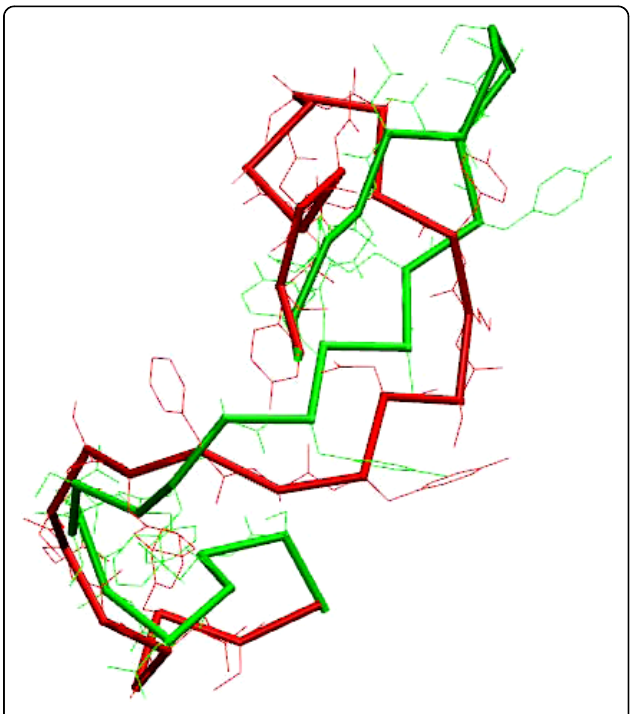


**Figure 6 Superimposition**. Loop fragment 106-130 of 1prn crystallographic structure (red) superimposed with CABS (cRMSD 4.80 Å) model (green).

Jamroz and Kolinski *BMC Structural Biology* 2010, **10**:5
http://www.biomedcentral.com/1472-6807/10/5


## Conclusions

In this work we have shown that *de novo* protein loops modeling using ROSETTA and CABS-based software is complementary to the classical modeling with MODELLER, the golden standard of comparative modeling. The proposed hybrid modeling pipeline, where ten top ranked (according to DOPE statistical potential) MODELLER models are used as templates for CABS, allows for meaningful loop modeling for a broad range of loop length. The hybrid MODELLER-CABS method takes advantage of the local accuracy of MODELLER structures and the efficient sampling of local free-energy minima by CABS. The hybrid-CABS method described in this work extends the applicability range of protein comparative modeling. Further increase of accuracy for large loops will require better ranking of resulting models. Model-ranking in the range of moderate- and low-resolution computational structures remains a challenging problem for the entire structure-prediction field. In this case, a small step in this direction was performed by a combination of different modeling techniques.

## Methods

The dataset employed in this work is summarized in Table 1. The cases of shorter loops, up to 12 residues, are taken from the work of Rossi et al. who used a loop database developed by Jacobson et al. [13,14]. The longer loops were selected from the same protein structures as continuous fragments of coil structures, according to the DSSP definition of secondary structure. Dangling ends are excluded from our test, similarly as it was done by others. Dangling ends are frequently structurally poorly defined, and therefore the results of their simulations are difficult to interpret. The dataset is available for download (Additional file 2).

### Loop modeling with MODELLER and ROSETTA

All loops were first modeled using MODELLER, version 9v5, and the model-loop procedure [1]. The 500 resulting models were ranked using DOPE statistical potentials. Subsequently, loop modeling was repeated using ROSETTA software, leading to 500 independent models, ranked by the ROSETTA force field [25]. The description of the CABS modeling tool and the procedure employed in present study is provided below.

### CABS modeling software

CABS is a versatile modeling tool, based on the coarse graining of polypeptide conformational space and knowledge-based force field. Applications of CABS include protein structure prediction (from comparative to template-free modeling), prediction of protein folding mechanisms and flexible modeling of macromolecular assemblies [3,23,24,26]. Technical details of CABS design and software are provided elsewhere [27]. At this point, for the reader's convenience, we provide only an outline of the most essential features. The CABS (C-alpha, C-beta, and Side chain) representation of protein conformational space employs a united residue approach. A single amino acid is represented by four pseudo-atoms: centered on the alpha carbon, on the beta carbon, in the center of mass of the side chain (where applicable) and an additional pseudo-atom located in the center of the virtual $C\alpha$-$C\alpha$ bond. The $C\alpha$ pseudo atoms are restricted to vertices of regular cubic lattice with the lattice spacing equal to 0.61 Å. Due to allowed fluctuations of the $C\alpha$-$C\alpha$ distance around the canonical value of 3.78 Å the set of possible representations of this virtual bond consists of 800 lattice vectors. Thus, serious lattice artifacts could be safely ignored. The accuracy of the $C\alpha$-trace projection onto this lattice is in the order of 0.35 Å. On the other hand, lattice representation smoothens the model energy landscape and speeds -up computation by using pre-computed local conformational transitions which require simple references to hashing tables instead of computing trigonometric transformations, as would be necessary in an otherwise equivalent continuous space model. Coordinates of other pseudo-atoms are off-lattice and are defined in the reference frame provided by the $C\alpha$ trace. Again, these coordinates are pre-computed and stored in simple reference tables, in which the two indices (range of 1-800, each) encode the conformation of three consecutive alpha carbons. It is assumed that coordinates of such fragments define positions of the side chain for the central residue.

Conformational updates include various local transformations, controlled by a pseudo random mechanism. There are single $C\alpha$ moves, two, three and four $C\alpha$ fragment transitions and small displacements of larger (4-22 residue) fragments. Update of a single $C\alpha$ position involves side chain updates of the central and two neighboring residues. The sampling scheme could be executed within a classical Metropolis Monte Carlo scheme (when isothermal dynamics is required) or using a Replica Exchange (REMC) protocol when equilibrium data are required only, as in the case of the present work.

The force field of CABS consist of several types of potentials, including the hard-core excluded volume of the main chain and $C\beta$ atoms, generic (sequence independent) short-range protein-like biases, making the model chain behaving like a generic polypeptide chain, sequence-dependent short-range statistical potential, context-specific pairwise interactions of the side chain united atoms, with repulsive and attractive square-well potential, and finally, a model of main-chain cooperative hydrogen bond networks. The details of the force field

could be found in earlier publications and the numerical data for the histogram-type potentials are available from the authors' homepage http://biocomp.chem.uw.edu.pl.

CABS allows for very straightforward implementation of restraints of various types. These may include soft biases towards predicted secondary structures and theoretically predicted side chain contacts, distance restraints read from templates for comparative modeling, restraints derived from sparse NMR data, etc.

### Loop modeling procedure with CABS

First, the template proteins (with excised loop fragments) are projected onto the CABS lattice, and the loop fragments are added in a random fashion. Then, the non-loop fragments of the original structures are used to read several hundreds of distance restraints, similarly to the procedures used in comparative modeling with CABS [3]. Subsequently, the starting structure is copied to 20 identical replicas for REMC simulations. During the REMC simulations temperatures of all replicas were gradually lowered, with a constant temperature distance between the replicas. Only the snapshots from the lowest temperature replica were stored in a pseudo-trajectory. Each simulation was repeated three times (with different streams of pseudo-random numbers), and the collated results were subject to final analysis. Trajectories were clustered using the K-means method. Also the medoids and the best observed structures from each trajectory were stored. It was observed that the centroids of the largest clusters were very close to the centroids from the entire trajectories. Thus, the trajectory medoid structures were reported as the top ranking models. For the top CABS structures the full atom molecular models were built using BBQ and SCWRL software [28,29]. Such a multiscale modeling strategy (from coarse-grained to all atom structures) proved very efficient in earlier applications of CABS software. Modeling of a single protein from the test set employed in this work using CABS protocol requires 10-15 hours of single LINUX box, which is similar to the cost of generating 500 structured by the ROSETTA method. Generation of 500 examples using MODELLER is 2-3 times faster.

### Hierarchical modeling with MODELLER and CABS

Analysis of preliminary modeling results led to an interesting observation: The distribution of the accuracy of the models generated by MODELLER and ROSETTA was significantly broader than the distribution of the quality of models generated by CABS.

The reason is that the models generated by MODELLER and ROSETTA are independent of one another, while the models from CABS are highly correlated along the simulation pseudo-trajectory. Consequently, the best models (among the 500 generated by MODELLER or ROSETTA) are usually considerably better than the top-ranked models. Unfortunately, the selection of the best models from a large set of decoys remains an unsolved problem for each of these methods. Taking the above into consideration, we designed a hybrid modeling pipeline that should take advantages of these methods. Namely, top ranked models from MODELLER (top 10) were used as structural templates for the derivation of distance restrains (including loop fragments) for modeling with CABS. It was expected that better local geometry of MODELLER structures and their diversity should improve sampling with CABS. The result of such an approach are reported as the CABS-hybrid simulations. Medoids (structures closest to the structural centroid from a pseudo-trajectory) were reported as the top ranked models. A similar modeling strategy was designed for a combination of ROSETTA and CABS. The accuracy of such an approach is similar to the accuracy of the aformentioned MODELLER-CABS hybrid. Since MODELLER is computationally less expensive than ROSETTA we present a benchmark only for the later combination.

**Additional file 1: Student t-test**. Results from two sample paired t-test of Table 2.
Click here for file
[ http://www.biomedcentral.com/content/supplementary/1472-6807-10-5-S1.GZ ]

**Additional file 2: Loop benchmark test set**. Database of 186 experimentally derived protein loop models used in the simulations.
Click here for file
[ http://www.biomedcentral.com/content/supplementary/1472-6807-10-5-S2.XLS ]

#### Acknowledgements

This work was supported by NIH grant no. 1R01GM081680 and Polish Ministry of Science and Higher Education grant no. NN301465634. The computational part of this work was executed using the computer cluster of the Computing Center of Faculty of Chemistry, University of Warsaw. MJ acknowledges the support from a Project operated within the Foundation for Polish Science MPD Programme co-financed by the EU European Regional Development Fund. A commercial version of CABS-based modeling software was used http://www.selvita.com/selvita-protein-modeling-platform.html.

#### Authors' contributions

AK conceived the use of a combination of modeling techniques. MJ performed simulations and analysis of the results. AK drafted the manuscript and both authors read and approved the final version of the manuscript.

Received: 8 October 2009
Accepted: 11 February 2010 Published: 11 February 2010

#### References

1. Eswar N, Eramian D, Webb B, Shen MY, Sali A: **Protein structure modeling with MODELLER.** *Methods Mol Biol* 2008, **426**:145-159.
2. Ginalski K: **Comparative modeling for protein structure prediction.** *Curr Opin Struct Biol* 2006, **16**:172-177.

3. Kolinski A, Bujnicki JM: **Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models.** *Proteins* 2005, **61**:84-90.
4. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A: **Critical assessment of methods of protein structure prediction-Round VII.** *Proteins* 2007, **69**:3-9.
5. Pokarowski P, Kloczkowski A, Jernigan RL, Kothari NS, Pokarowska M, Kolinski A: **Inferring ideal amino acid interaction forms from statistical protein contact potentials.** *Proteins* 2005, **59**:49-57.
6. Rohl CA, Strauss CEM, Misura KMS, Baker D: **Protein Structure Prediction Using Rosetta.** *Numerical Computer Methods, Part D, of Methods Enzymol* Academic PressBrand L, Johnson ML 2004, **383**:66-93.
7. Friedberg I: **The interplay of fold recognition and experimental structure determination in structural genomics.** *Curr Opin Struct Biol* 2004, **14**:307-312.
8. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** *J Mol Biol* 1993, **234**:779-815.
9. Baker D, Sali A: **Protein structure prediction and structural genomics.** *Science* 2001, **294**:93-96.
10. Kmiecik S, Gront D, Kolinski A: **Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field.** *BMC Struct Biol* 2007, **7**:43.
11. Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM: **A "FRankenstein's monster" approach to comparative modeling: Merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation.** *Proteins* 2003, **53**:369-379.
12. Kolinski A, Gront D: **Comparative modeling without implicit sequence alignments.** *Bioinformatics* 2007, **23**:2522-2527.
13. Rossi KA, Weigelt CA, Nayeem A, Krystek SR: **Loopholes and missing links in protein modeling.** *Protein Sci* 2007, **16**:1999-2012.
14. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA: **A hierarchical approach to all-atom protein loop prediction.** *Proteins* 2004, **55**:351-367.
15. Hildebrand PWW, Goede A, Bauer RAA, Gruening B, Ismer J, Michalsky E, Preissner R: **SuperLooper-a prediction server for the modeling of loops in globular and membrane proteins.** *Nucleic Acids Res* 2009, **37**:W571-W574.
16. Spassov VZ, Flook PK, Yan L: **LOOPER: a molecular mechanics-based algorithm for protein loop prediction.** *Protein Engineering, Design and Selection* 2008, **21**:91-100.
17. Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJ, Oliva B: **ArchDB: automated protein loop classification as a tool for structural genomics.** *Nucl Acids Res* 2004, **32**:D185-188.
18. Peng HP, Yang AS: **Modeling protein loops with knowledge-based prediction of sequence-structure alignment.** *Bioinformatics* 2007, **23(Suppl 7)**:2836-2842.
19. Fernandez-Fuentes N, Zhai J, Fiser A: **ArchPRED: a template based loop structure prediction server.** *Nucl Acids Res* 2006, **34**:W173-176.
20. Soto CSS, Fasnacht M, Zhu J, Forrest L, Honig B: **Loop modeling: sampling, filtering, and scoring.** *Proteins* 2008, **70**:834-843.
21. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A: **Protein fragment reconstruction using various modeling techniques.** *J Comput Aided Mol Des* 2003, **17**:725-738.
22. Shen MY, Sali A: **Statistical potential for assessment and prediction of protein structures.** *Protein Sci* 2006, **15**:2507-2524.
23. Kmiecik S, Kolinski A: **Characterization of protein-folding pathways by reduced-space modeling.** *Proc Natl Acad Sci USA* 2007, **104**:12330-12335.
24. Kmiecik S, Kolinski A: **Folding pathway of the B1 domain of protein G explored by a multiscale modeling.** *Biophys J* 2008, **94**:726-736.
25. Wang C, Bradley P, Baker D: **Protein-Protein Docking with Backbone Flexibility.** *J Mol Biol* 2007, **373**:503-519.
26. Kurcinski M, Kolinski A: **Hierarchical modeling of protein interactions.** *J Mol Model* 2007, **13**:691-698.
27. Kolinski A: **Protein modeling and structure prediction with a reduced representation.** *Acta Biochim Pol* 2004, **51**:349-371.
28. Gront D, Kmiecik S, Kolinski A: **Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates.** *J Comput Chem* 2007, **28**:1593-1597.
29. Canutescu AA, Shelenkov AA, Dunbrack RL: **A graph-theory algorithm for rapid protein side-chain prediction.** *Protein Sci* 2003, **12**:2001-2014.

doi:10.1186/1472-6807-10-5
**Cite this article as:** Jamroz and Kolinski: **Modeling of loops in proteins: a multi-method approach.** *BMC Structural Biology* 2010 **10**:5.

# B | DESIGNING AN AUTOMATIC PIPELINE FOR PROTEIN STRUCTURE PREDICTION

*Kmiecik S, Jamroz M, Zwolinska A, Gniewek P, Kolinski A. (2008)*
Designing an Automatic Pipeline for Protein Structure Prediction. *In: Hansmann U, Meinke J, Mohanty S, Nadler W, Zimmermann O, eds. From Computational Biophysics to Systems Biology (CBSB08) Proceedings, NIC Series Vol. 40.Vol 40. Jülich; 105–108.*

# Designing an Automatic Pipeline for Protein Structure Prediction

**Sebastian Kmiecik**[1], **Michal Jamroz**[1], **Anna Zwolinska**[1], **Pawel Gniewek**[1], **and Andrzej Kolinski**[2]

[1] Selvita Sp. z o.o.,
Ostatnia 1c, 31-444 Cracow, Poland
*E-mail: sebastian.kmiecik@selvita.com*

[2] Laboratory of Theory of Biopolymers,
Faculty of Chemistry, University of Warsaw,
02-093 Warsaw, Poland

Building accurate 3D structural models of proteins and protein assemblies is a challenging task. Our modeling technology is based on the CABS model, extensively tested, state-of-the-art approach to protein structure prediction. The modeling process is divided into two stages: CABS fold assembly followed by the model refinement/selection procedure, using an all-atom representation and a more exact interaction scheme enabling high resolution structure prediction. Fold assembly can be done in a framework of a standard comparative modeling procedure, where spatial restraints are derived from alternative sequence alignments with a template/templates. Preferentially in more difficult modeling cases, a new approach to comparative modeling can be used, which does not require the prior alignment. Selvita's goal is to provide an integrated tool-kit for automated protein structure predictions. However, like blind prediction experiments show, due to high complexity of prediction tasks, fully automated approach often doesn't guarantee the highest possible performance. Therefore, human intervention is made possible at every stage of modeling.

## 1 Introduction

Thanks to international effort in the genome sequencing projects, enormous library of protein sequences is now available. Despite extensive efforts in structural genomics, the number of experimentally determined protein structures, typically by costly X-ray crystallography or NMR spectroscopy procedures, is lagging far behind the number of known protein sequences. Since proteins are involved in practically all functions performed by a cell, knowledge of protein structures is necessary for understanding and controlling molecular mechanisms of life. Current assumptions are, that for a large fraction of proteins whose structures will not be determined experimentally, computational methods can provide valuable information[1].

## 2 Multiscale Approach to Structure Prediction: Comparative Modeling and Fold Recognition

During computational protein structure determination the following main challenges can be identified: 1) High accuracy structure prediction, at the resolution comparable to experimental methods, to enable predicted models utilization in a number of protein structure-based approaches (e.g. drug design, protein design, molecular docking, molecular replacement), which is now possible in Comparative Modeling (CM) cases[2], 2) Structure prediction of proteins or protein fragments for which sequence search methods failed to find

unambiguous homologs with known structure (Fold Recognition (FR) and New Fold (NF) prediction)

To meet criteria of both challenges, precise interaction scheme, sensitive to small atomic rearrangement, should be somehow combined with high efficiency in exploring proteins conformational space. That can be achieved by combining all-atom and reduced modeling: the multiscale modeling. Properly designed reduced models make possible very effective search of the protein's conformational space[3] and all-atom modeling enable exact scoring and refinement of the models. Our modeling technology is based on a such hierarchical approach[2]. Reduced-space search of the conformational space by the CABS[3] is followed by a reliable transition into the all-atom resolution and by subsequent fine-tuning and assessment of the final models. Such multiscale approach enable high-resolution protein structure predictions, predictions of protein interactions[4], computer-aided drug design and even study of protein dynamics[5].

CABS computational technology has been rigorously tested during CASP6 (Critical Assessment of Techniques for Protein Structure Prediction) world-wide experiment by the Kolinski-Bujnicki group, which ranked second best among over 200 groups participating, and ranked first when the consistency of the prediction was used as a criterion (the number of CASP targets placed in the top 20 of the best predictions)[6].

The design of CABS model enable easy implementation of spatial restraints. Such restraints can be derived by a large number of bioinformatics tools from appropriate known structures or from experimental sources e.g. from sparse NMR data. Therefore, essentially the same approach is possible at various levels of protein modeling difficulty from CM, to FR and NF cases. For the sake of flexibility two basic modeling pathways were designed and one alternative to make the prediction more effective. The entire prediction pipeline could by briefly outlined as follows (see the flowchart in the Figure 1): 1) Pre-processing: Template identification, secondary structure prediction, target- template alignments, input for more sophisticated user defined FR multiple alignments, 2a) Fast modeling track (easy CM cases) including fast scoring of alternative alignments and generation of spatial restraints, 2b) Rigorous modeling track (hard CM and FR cases) including 3D threading and generation of spatial restraints, 2c) Alternative modeling track by TRACER (hard CM and FR cases) - without prior alignments[7], 3) CABS modeling, 4) Post-processing: trajectory clustering, selection of clusters representatives, rebuilding from reduced to all-atom representation and finally all-atom models refinement and ranking.

Additionally in the most difficult cases (NF) ab initio modeling based only on target sequence can be performed (the accuracy of the resulting models is sometimes sufficient for structure-based protein function identification).

## 3 Automatic or Human Driven?

As blind structure prediction experiments demonstrated, human expert experience and intuition becomes a key point to the best possible performance, especially in difficult CM and FR[1]. Also in high resolution structure prediction, when a fraction of an Angstrom of the final model resolution matters, human intervention may by helpful by manual insertions of a template structure fragments into the final model. However, our goal is to develop fully automated structure prediction protocol which enable structure prediction on a genomic scale. Considering difficult modeling cases, the modeling approach without prior align-

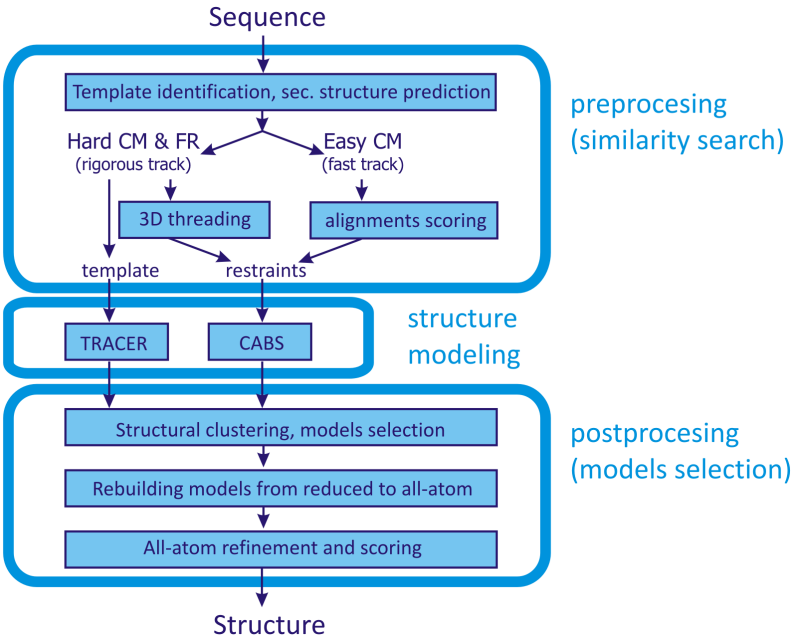



Figure 1. The protein structure prediction flowchart - see the text.

ments[7], included in our pipeline, seems to be an extremely promising step towards fully automated modeling (errors in alignments seem to be the main source of failures in protein structure prediction[1]).

## References


1. O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, D. Baker, *Progress in modeling of protein structures and interactions*, Science **310**, 638-42, 2005.
2. S. Kmiecik, D. Gront, A. Kolinski, *Towards high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field.*, BMC Structural Biology , 7:43, 2007.
3. A. Kolinski, *Protein modeling and structure prediction with a reduced representation*, Acta Biochim. Pol. **51**, 349-371, 2004.
4. M. Kurcinski, A. Kolinski, *Hierarchical modeling of protein interactions*, J Mol Model **13**, 691-8, 2007.
5. DA Debe, JF Danzer, WA Goddard, A. Poleksic, *STRUCTFAST: Protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring*, Proteins **64**, 960-967, 2006.
6. S. Kmiecik, A. Kolinski, *Characterization of protein-folding pathways by reduced-space modeling.*, Proc Natl Acad Sci USA **104**, 12330-5, 2007.
7. A. Kolinski, D. Gront, *Comparative modeling without implicit sequence alignments.*, Bioinformatics **23**, 2522-27, 2007.

# C | PROTEIN STRUCTURE PREDICTION USING CABS - A CONSENSUS APPROACH

*Blaszczyk M, Jamroz M, Gront D, Kolinski A. (2012)* Protein Structure Prediction Using CABS – A Consensus Approach. *In: Carloni P, Hansmann U, Lippert T, et al., eds. From Computational Biophysics to Systems Biology (CBSB11) Proceedings, IAS Series: Vol. 8. Jülich; 29–32.*

# Protein Structure Prediction Using CABS – A Consensus Approach

**Maciej Blaszczyk, Michal Jamroz, Dominik Gront, and Andrzej Kolinski**

Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw
02-093 Warsaw, Poland
*E-mail: mblaszczyk@chem.uw.edu.pl*

We have designed a new pipeline for protein structure prediction based on the CABS engine. The procedure is fully automated and generates consensus models from a set of templates. Restraints derived from the templates define a region of conformational space, which is then sampled by Replica Exchange Monte Carlo algorithm implemented in CABS. Results from CASP9 show, that for great majority of targets this approach leads to better models than the mean quality of templates (in respect to GDT_TS). In five cases the obtained models were the best among all predictions submitted to CASP9 as the first models.

## 1 Introduction

Knowledge of 3D structures of proteins is a crucial requirement for a progress in many areas of biomedicine, e.g. rational drug design. Due to the complexity and high cost of structure determination by experimental methods (mainly Xray crystallography or NMR), computer-based protein structure prediction methods have been placed in the center of attention of a broad community of molecular and cell biologists[1]. Nowadays, there is a number of publicly available web servers, which provide methods for protein structure prediction[2]. Moreover, thanks to the meta-servers[3,4], which collect data from servers, obtaining the predictions is even easier. However, for most purposes it is necessary to provide one, possibly the best, final model. A common approach to this problem is the use of Model Quality Assessment Programs (MQAPs) which score models according to various criteria[5] and selection of the top scoring one. Obviously, the MQAPs can't propose a model better then the best of input structures. Application of CABS modeling tool[6] with spatial restraints derived from the templates allows for reaching beyond this limit.

## 2 Methods

The procedure used during CASP9 consisted of several steps (Fig. 1) and was trained on the targets from previous CASPs. The first step was templates selection. As templates we used server predictions submitted to CASP9. The list of the servers from which models were taken, was created on the basis of their performance during the CASP8. To check if the best servers from CASP8 are still the reliable ones, servers predictions from CASP9 were ranked using 3D-jury score[7]. Then, for all selected templates distances between pairs of alpha carbons were extracted[8]. The minimum and the maximum distance between pairs of residues were taken as limits of the ranges of restraints. Using templates as a starting structures we have run two independent Replica Exchange Monte Carlo simulations with CABS[6].

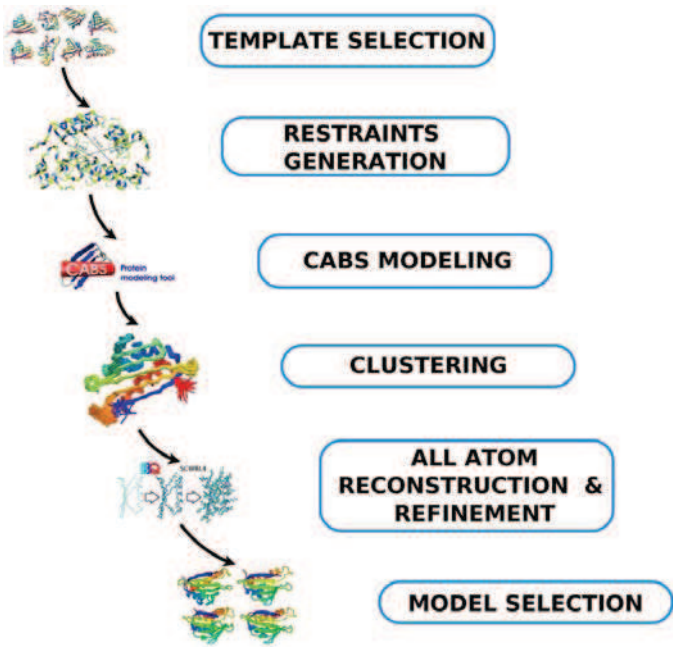



Figure 1. Flowchart of the pipeline used during CASP9. See the text for details.

CABS is a lattice model with a representation reduced to four united atoms per residue: C$\alpha$, C$\beta$, center of mass of a side chain (where applicable) and the center of a virtual C$\alpha$ – C$\alpha$ bond. The force field of the model employs knowledge based potentials derived from the statistical analysis of the databases containing known protein structures. Conformational space is sampled using Replica Exchange Monte Carlo method. Application of the restraints reduces conformational space for sampling, which makes modeling faster and more accurate.

The resulted trajectories from CABS were clustered[9], and the clusters' centroids were calculated. Because of reduced representation in CABS, it was necessary to rebuilt the atomistic details of obtained models. Reconstruction of the backbone using BBQ[10] was followed by reconstruction of the side chains with SCWRL4[11]. Next, we performed model refinement, which was also done in two steps. To improve model geometry (e.g. bond length) we employed Modeller[12]. Then, we used GROMACS[13] in order to refine some packing details. Finally, obtained models were ranked on the basis of the clusters' density and the level of similarity of the models from two independent simulations.

## 3 Results

Since the presented method aims at a consensus prediction from a set of templates it is worth to compare the accuracy of obtained models and the templates used. For great majority of targets GDT_TS of the model was higher then mean GDT_TS of templates. Moreover, in 5 cases the accuracy of the model was better then the accuracy of the best template (see Fig. 2).

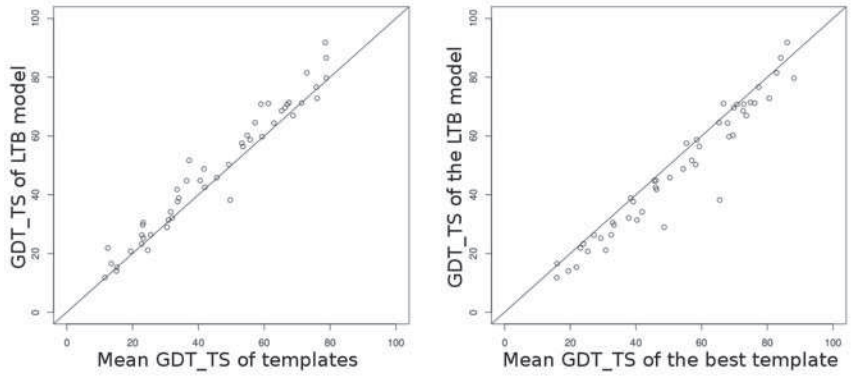



Figure 2. Comparison of GDT_TS of templates and obtained models.

According to the official assessment our models (from Laboratory of Theory of Biopolymers - LTB) for 5 selected domains were the best among all predictions submitted to CASP9 as the first models. As shown in Fig. 3, for great majority of targets, GDT_TS of obtained structure was higher then mean GDT_TS of all models submitted to the CASP. However, there are a few cases with significant losses of accuracy . Most of them are large multi-domain proteins, for which it was necessary to perform domain division, which was not supported in the procedure. This problem is to be solved in a future work.

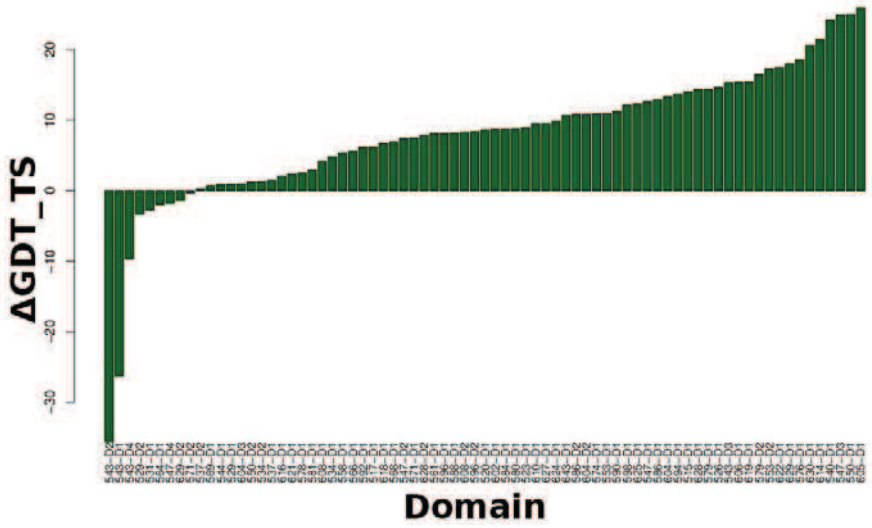



Figure 3. Differences between GDT_TS scores of our models and the mean for all models submitted to CASP.

## Acknowledgments

Support from Marie Curie fellowship (FP7-people-IOF) for DG is acknowledged. Computational part of this work was done using the computer cluster at the Computing Center of Faculty of Chemistry, University of Warsaw.

## References

1. Yang Zhang, *Protein structure prediction: when is it useful?*, Current opinion in structural biology, **19**, no. 2, 145–155, Apr. 2009.
2. Daniel Fischer, *Servers for protein structure prediction.*, Current opinion in structural biology, Mar. 2006.
3. Krzysztof Ginalski, Arne Elofsson, Daniel Fischer, and Leszek Rychlewski, *3D-Jury: a simple approach to improve protein structure predictions*, Bioinformatics, **19**, no. 8, 1015–1018, May 2003.
4. Jesper Lundström, Leszek Rychlewski, Janusz Bujnicki, and Arne Elofsson, *Pcons: a neural-network-based consensus predictor that improves fold recognition.*, Protein science : a publication of the Protein Society, **10**, no. 11, 2354–2362, Nov. 2001.
5. Andriy Kryshtafovych and Krzysztof Fidelis, *Protein structure prediction and model quality assessment.*, Drug discovery today, **14**, no. 7-8, 386–393, Apr. 2009.
6. Andrzej Kolinski, *Protein modeling and structure prediction with a reduced representation.*, Acta biochimica Polonica, **51**, no. 2, 349–371, 2004.
7. László Kaján and Leszek Rychlewski, *Evaluation of 3D-Jury on CASP7 models*, BMC Bioinformatics, **8**, 304+, Aug. 2007.
8. Dominik Gront and Andrzej Kolinski, *Utility library for structural bioinformatics*, Bioinformatics, **24**, no. 4, 584–585, Feb. 2008.
9. Dominik Gront and Andrzej Kolinski, *HCPM–program for hierarchical clustering of protein models.*, Bioinformatics, **21**, no. 14, 3179–3180, July 2005.
10. Dominik Gront, Sebastian Kmiecik, and Andrzej Kolinski, *Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates*, J. Comput. Chem., **28**, no. 9, 1593–1597, July 2007.
11. Adrian A. Canutescu, Andrew A. Shelenkov, and Roland L. Dunbrack, *A graph-theory algorithm for rapid protein side-chain prediction*, Protein Science, **12**, no. 9, 2001–2014, Sept. 2003.
12. Narayanan Eswar, Ben Webb, Marc A. Marti-Renom, M. S. Madhusudhan, David Eramian, Min-yi Shen, Ursula Pieper, and Andrej Sali, *Comparative Protein Structure Modeling Using Modeller*, Current protocols in bioinformatics, **Chapter 5**, Oct. 2002.
13. David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. Berendsen, *GROMACS: fast, flexible, and free.*, Journal of computational chemistry, **26**, no. 16, 1701–1718, Dec. 2005.

# D | STRUCTURAL FEATURES THAT PREDICT REAL-VALUE FLUCTUATIONS OF GLOBULAR PROTEINS

*Jamroz M, Kolinski A, Kihara D. (2012)* Structural features that predict real-value fluctuations of globular proteins. *Proteins 80(5):1425–35.*

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# Structural features that predict real-value fluctuations of globular proteins

**Michal Jamroz,[1,2] Andrzej Kolinski,[1] and Daisuke Kihara[2,3,4]***

[1] Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warszawa, Poland
[2] Department of Biological Sciences, College of Science, Purdue University, West Lafayette, Indiana 47907
[3] Department of Computer Science, College of Science, Purdue University, West Lafayette, Indiana 47907
[4] Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, Indiana 47907

## ABSTRACT

**It is crucial to consider dynamics for understanding the biological function of proteins. We used a large number of molecular dynamics (MD) trajectories of nonhomologous proteins as references and examined static structural features of proteins that are most relevant to fluctuations. We examined correlation of individual structural features with fluctuations and further investigated effective combinations of features for predicting the real value of residue fluctuations using the support vector regression (SVR). It was found that some structural features have higher correlation than crystallographic *B*-factors with fluctuations observed in MD trajectories. Moreover, SVR that uses combinations of static structural features showed accurate prediction of fluctuations with an average Pearson's correlation coefficient of 0.669 and a root mean square error of 1.04 Å. This correlation coefficient is higher than the one observed in predictions by the Gaussian network model (GNM). An advantage of the developed method over the GNMs is that the former predicts the real value of fluctuation. The results help improve our understanding of relationships between protein structure and fluctuation. Furthermore, the developed method provides a convienient practial way to predict fluctuations of proteins using easily computed static structural features of proteins.**

Proteins 2012; 00:000–000.


**Key words: protein flexibility; protein dynamics; structure-dynamics relationship; support vector regression; molecular dynamics; fluctuation prediction.**

## INTRODUCTION

Thanks to worldwide efforts in structural genomics,[1–3] we now know over 75,000 protein tertiary structures.[4] This number is only a small fraction when compared with the number of known protein sequences. Computational methods can predict structures for more than a half of newly sequenced proteins by means of template-based modeling with a sufficiently high accuracy.[5–8] For some of the remaining proteins, it is possible to predict their structures in a de novo fashion if they are small and structurally simple.[9–14] Thus, the problem of protein structure prediction is practically gradually being solved, and it may be completely solved in the near future. Obviously, for the most difficult (and "atypical") cases of monomeric structures and to a much larger extent for the plethora of possible protein–protein (protein-nucleic acid, protein-carbohydrate, etc.) complexes, structure prediction will remain a challenging task for decades.[9,15–17] The knowledge of protein tertiary structures facilitates fast developments in various branches of molecular medicine and biotechnology.[18,19] It, however, becomes more and more obvious that to understand the underlying molecular mechanisms of life, we need to see biomolecules "in action."

Protein dynamics, resulting from a specific flexibility of their structures, has drawn much attention recently in both theoretical and experimental molecular biology. Studies of dynamics of protein structures and their assemblies are important for understanding the mechanisms of protein function in various cellular processes,[20,21] in particular, ligand binding, enzymatic reactions,[22] conformational diseases,[23] and protein–protein interaction.[24] The understanding of protein flexibility is also important for practical applications such as development of computer-aided methods of enzyme design[25,26] and drug development.[27]

In X-ray protein crystallography, which determines the Cartesian coordinates of atoms in proteins, uncertainties/fluctuations of atomic positions are provided in the form of *B*-factors.[28] The *B*-factor measures the mobility of atoms, but it also reflects some inherent aspects of crystallographic

Grant sponsor: EU European Regional Development Fund (Foundation for Polish Science MPD Programme); Grant sponsor: National Science Foundation; Grant number: IIS0915801; Grant sponsor: National Institutes of Health; Grant numbers: R01GM075004, R01GM097528; Grant sponsor: National Science Foundation; Grant numbers: DMS0800568, EF0850009
*Correspondence to: Daisuke Kihara, Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN 47907. E-mail: dkihara@purdue.edu.
Received 2 December 2011; Revised 3 January 2012; Accepted 11 January 2012
Published online 27 January 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24040

techniques. Moreover, fluctuations estimated by *B*-factors are influenced by the molecular environment of the crystal structure. Protein mobility in solution could differ qualitatively from that in a crystal. Eastman *et al.*[29] showed that *B*-factors are an accurate measure of fluctuations for stable parts of proteins, but significantly underestimate motion in flexible regions. Somewhat more straightforward measures of structure fluctuations could be derived from nucleic magnetic resonance (NMR) experiments, although resulting estimates can be flawed by various limitations of actual measurements and by the computational schemes of their interpretation.[30–33] Therefore, these methods do not fully reflect actual fluctuations of proteins.

Molecular dynamics (MD) is the most straightforward method for theoretical studies of dynamic aspects of molecular systems. Because of the progress in computing technology, it is now practical to simulate protein systems in a timescale of tens of nanoseconds. Nevertheless, such simulations remain costly. With a significantly less computational requirement, the internal motion of a protein can be approximated by the normal mode analysis of a harmonic model of proteins.[34] Another possibility is to use simulations using coarse-grained representations of protein structures. A simple approach is the Gaussian Network Model (GNM) and its derivatives.[35–38] Long-time simulation at an intermediate resolution can be achieved using simplified protein models such as UNRES[39] and CABS.[40] These models enable a low-resolution study of dynamics (or stochastic dynamics) in timescales by a few orders of magnitude longer than possible by all-atom MD.[41–44] A weak point of studying dynamics using coarse-grained models is a lack of straightforward scaling between the models' time and the real time. Thus, all-atom MD simulations should always be used as a reference for coarse-grained dynamics.

A number of computational methods for predicting protein fluctuations have been published; however, almost all of them evaluated their prediction results mainly in comparison with the crystallographic *B*-factor of proteins. As discussed earlier, the *B*-factor does not fully capture the mobility of proteins in solution. As we show in this work, the fluctuations observed in MD and the *B*-factor correlate rather poorly, as was also concluded in a previous work.[29]

There are a series of works that use GNM or its variants for predicting *B*-factors of proteins.[35,38,45,46] Micheletti *et al.*[47] extended GNM by adding Cβ atoms (βGM). The fluctuations of residues predicted by βGM were compared to the fluctuations from the MD simulation of HIV-1 protease. The self-consistent pair contact probability method, which is similar in its spirit to GNM, was used to predict fluctuations and compared with *B*-factors.[48] Zhou and coworkers[49] developed an all-atom mean-field model to predict fluctuations.

Structural features of proteins were also investigated that can indicate fluctuations represented by *B*-factors. These features include solvent accessibility of residues,[50] distance from a residue to the center of mass of the protein,[51] eigenvectors of the square distance matrix,[52] and predicted local fragment structures.[53] An alternative direction pursued was to predict *B*-factors from protein sequences. Machine-learning methods, such as Support Vector Machine,[54,55] the random forest algorithm,[56] or an artificial neural network,[57] were used to predict fluctuations using sequence information and structural features that can be predicted from sequences, such as the secondary structure and the accessible surface area of residues.

In this work, we used support vector regression (SVR) to investigate the relationship between protein structure and dynamics. We used various structural characteristics as well as structure fluctuation profiles predicted by GNM as input for SVR. The target reference is the dynamics observed in long MD simulations for a representative set of 592 globular proteins. To the best of our knowledge, this is the first time that protein fluctuations have been investigated on such a large dataset of MD simulations. In this context, we also analyzed differences of protein dynamics deducted from the *B*-factors and the in-solvent dynamics computed by MD simulations. A more practical purpose of this work is to provide a fast (essentially instantaneous in comparison with MD) and reliable method that can be used for predicting fluctuations of protein structures. Unlike existing works mentioned earlier, we predict the real value of residue fluctuations rather than simply showing correlation between predicted and actual fluctuations values. Remarkably, our method predicts fluctuation highly accurately with an average error of less than 1.1 Å. The correlation coefficient of our prediction with the actual fluctuations observed in MD simulations is higher than that of GNM. We also found that some of the static structural features, such as residue contact number, have higher correlation with the residue fluctuation in MD simulation than *B*-factors do. The developed software for predicting fluctuation, named flexPred, has been made freely available for the academic community.

## MATERIALS AND METHODS

### Dataset of molecular dynamics trajectories

The molecular dynamics (MD) trajectories of proteins were selected from MoDEL (Molecular Dynamics Extended Library).[58] Of 1897 entries in the database, the following entries were discarded: trajectories for protein structures solved by NMR, those which include more than one protein chain in the simulation, and trajectories for proteins whose length differ from the corresponding entries in the Protein Data Bank (PDB).[4] These MD

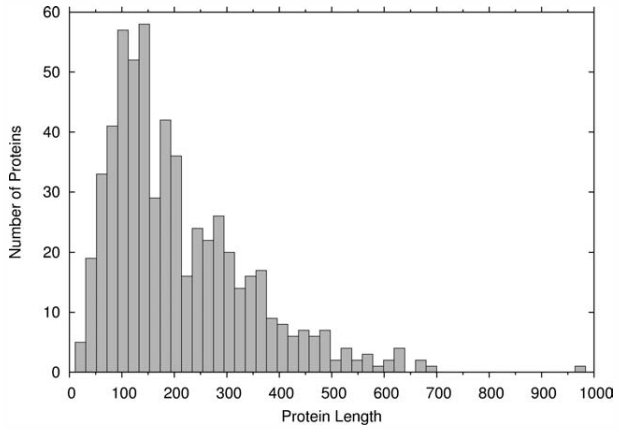



**Figure 1**

Histogram of the length of proteins in the dataset. There are in total 592 proteins.

trajectories were computed using AMBER,[59] GROMACS,[60] or NAMD[61] force fields. If more than one simulation is available for a protein, we used the first one with an earlier entry date in the database. The MoDEL trajectory files were uncompressed with the PCASuite software.[62] Eight hundred and thirty-seven trajectories remained after this filtering process. From this subset, we removed redundant proteins using the PISCES server[63] with a sequence identity cutoff of 35%. The final number of trajectories is 592. This dataset contains proteins from all main classes in the CATH database[64]: 111 proteins in the α class (18.75%), 149 proteins in the β class (25.17%), 256 in the αβ class (43.24%), and 76 in the few secondary structure class (12.84%). The length of the proteins ranges from 21 to 994 residues (Fig. 1). The simulation time was 10 ns for most of the proteins (96.11%), while the rest of the proteins had shorter trajectories: 5 (0.33%), 2 (2.36%), and 1 ns (0.5%), and one protein each with 6.5, 6.0, 5.5, and 4.5 ns.

## Definition of fluctuation

The fluctuation of amino acid residue $i$ is defined in two ways. It can be defined as a root mean square deviation (RMSD) of the mean position of an atom in an MD trajectory:

$$\sqrt{\left\langle (\Delta R_i)^2 \right\rangle}^{MD} = \sqrt{\frac{1}{T}\sum_{t_j=1}^{T}\left(x_i(t_j) - \left\langle x_i \right\rangle\right)^2} \qquad (1)$$

where $x_i(t_j)$ is the Cartesian coordinates of the Cα atom of residue $i$ at time $t_j$ in the trajectory, $T$ is the number of time frames in the trajectory, and $\langle x_i \rangle$ is the average position of the Cα atom of residue $i$ in the trajectory.

We also used the coordinates in the PDB file as the reference:

$$\sqrt{\left\langle (\Delta R_i)^2 \right\rangle}^{ref} = \sqrt{\frac{1}{T}\sum_{t_j=1}^{T}\left(x_i(t_j) - x_i^{ref}\right)^2}, \qquad (2)$$

where $x_i^{ref}$ is the coordinates of the Cα atom of residue $i$ in the PDB file. The distance of residue positions is computed after superimposing the PDB structure on each frame. If alternative positions of the atom are recorded in the PDB files, the first position of the atom was used. As shown in Figure 2, these two definitions give similar fluctuations of residues, but not identical. The correlation coefficient of the two fluctuation values is 0.86. The fluctuation value is smaller when the mean of a trajectory is used as the reference [Eq. (1)] in almost all the cases (99.9%). Unless noted, we use the second definition of fluctuation [Eq. (2)] in the results that will be shown below, because we compare the fluctuations from MD with *B*-factors and GNM, both of which are attributed to PDB structures.

## Structural features of proteins

We considered the following static protein structural features.

1. *B*-factor (temperature factor).[28] The *B*-factor reflects dynamic motion, the static disorder of the atom in the crystal structure, and also errors in model building. The *B*-factor values are taken from the PDB file.
2. Square of the distance between a residue and the protein center of mass, which is defined as follows:

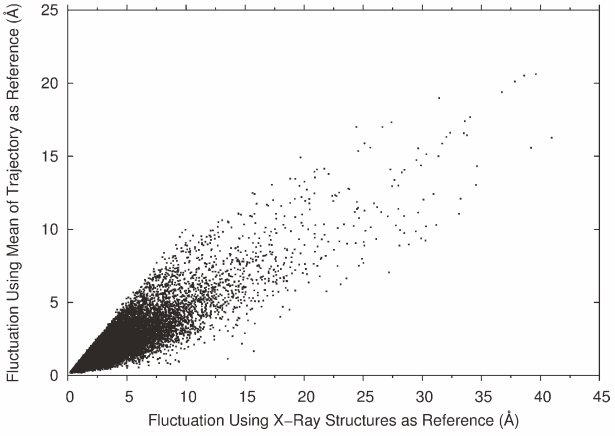



**Figure 2**

Average fluctuations of proteins in MD trajectories using two definitions. $x$ values show fluctuations of residues relative to the crystal structures of proteins in the PDB [Eq. (2)], while $y$ values are fluctuations relative to the mean structure of each MD trajectory [Eq. (1)].

$$r_i^2 = \left( x_i - 1/N \sum_{j=1}^{N} x_j \right)^2, \quad (3)$$

where $x_i$ is the position of the Cα atom of residue $i$. A previous work showed that this parameter has good correlation with the $B$-factor.[51,52]

3. Residue contact number, which is defined as the number of surrounding residues, whose Cα atom is closer than a cutoff distance. The contact number was also shown to correlate well with the $B$-factor.[65,66]
4. Number of hydrophobic/hydrophilic residue contacts, where the number of residue contacts is separately counted for hydrophobic and hydrophilic residues. Hydrophobic/hydrophilic residues are those which have a positive/negative value on the Kyte–Doolittle hydrophobicity scale.[67]
5. Solvent accessibility surface area (Å$^2$). This parameter is defined as water exposed surface of a residue. We used the DSSP program[68] to compute the accessibility surface area of amino acids, which are then normalized with the value in the tripeptide with glycines on both sides of the target amino acid residue.[69]
6. Residue depth, which is defined as the distance of the Cα atom or the average distance of all the atoms in a residue to the closest water molecule.[70] Protein surface was computed with the MSMS program.[71] The *hsexpo* program was used to compute residue depth.[72]
7. Lower/upper half-sphere exposure of a residue,[72] which is defined as the number of contacts within a half-sphere of a radius of 13 Å centering at either the Cα or the Cβ atom of the residue. The sphere is divided into half by a plane perpendicular to the Cα–Cβ vector.
8. Secondary structure. Each residue is classified into eight classes, that is, seven secondary structure types defined by DSSP[68] or other.
9. Fluctuations predicted by the GNM.[35,36] GNM is a coarse-grained model, where Cα atoms are connected by springs. GNM has been used for investigating protein dynamics including the prediction of $B$-factor values of proteins.[38] We downloaded GNM codes from the Jernigan laboratory (http://ribosome.bb.iastate.edu/). Fluctuations were computed with a residue contact distance cutoff of 16 Å[73] and without using cutoff.[38] Residue contacts in a protein are represented as the Kirchhoff matrix in GNM:

$$\Gamma_{ij} = \begin{cases} -1 & \text{if} \quad i \neq j \quad \text{and} \quad r_{ij} \leq r_c \\ 0 & \text{if} \quad i \neq j \quad \text{and} \quad r_{ij} \geq r_c \\ -\sum_{i,i\neq j}^{N} \Gamma_{ij} & \text{if} \quad i = j \end{cases}, \quad (4)$$

where $r_{ij}$ is the distance between two atoms, $i$ and $j$, and $r_c$ (=16 Å) is the cut-off value. GNM without cutoff uses the following modified Kirchhoff matrix:

$$\Gamma_{ij} = \begin{cases} r_{ij}^{-2} & \text{if} \quad i \neq j \\ -\sum_{i,i\neq j}^{N} \Gamma_{ij} & \text{if} \quad i = j \end{cases}. \quad (5)$$

In both methods, the average fluctuation of residue $i$ over time is defined by

$$\langle (\Delta R_i)^2 \rangle = C(\Gamma_{ii}^{-1}), \quad (6)$$

where $C$ is constant.

## Support vector regression

We combined the structural features listed above to predict fluctuations using support vector regression (SVR). The LIBSVM package[74] with Gaussian kernels was used. Because it was not feasible to test all the possible combinations of features, features were added or changed one at a time starting from the one which has the largest correlation coefficient with residue fluctuation. We performed fivefold cross validation using the dataset of trajectories. The default set of parameters in *libsvm*, $C = 64.0$, $\gamma = 1$, and $\epsilon = 0.5$, was used, which was shown to perform best among others tested in the first few feature combinations in the five-fold cross validation (data not shown).

## Evaluation of fluctuations prediction

Pearson's correlation coefficient was used to examine how well individual features or predicted fluctuations correlated with actual fluctuations in the MD trajectories. Average correlation coefficients were computed using all the trajectories in the dataset.

In addition, the error of predicted fluctuations was quantified as the RMSD to the reference trajectory fluctuation:

$$\mathrm{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \Delta R_i^{\mathrm{pred}} - \sqrt{\langle (\Delta R_i)^2 \rangle}^{\mathrm{ref}} \right)^2}, \quad (7)$$

where $N$ is the length of the protein, $\Delta R_i^{\mathrm{pred}}$ is predicted, and $\sqrt{\langle (\Delta R_i)^2 \rangle}^{\mathrm{ref}}$ is actual fluctuation [Eq. (2)] of residue $i$.

## Availability of the developed program

The program for predicting the fluctuation of residues in a protein structure is made freely available for the academic community at http://kiharalab.org/flexPred/. Both the web server and the source code written in Python are available. It takes a PDB file of a query protein for input data and outputs a predicted fluctuation value for each residue. The computational time for a protein is typically within a couple of seconds to 20 s depending on the length of the protein.

## RESULTS AND DISCUSSION

The relationships between structural features and residue fluctuations are examined in several aspects. First, we compare the correlation coefficient of individual static structural features with actual fluctuations. Then, we explore different combinations of features to make accurate prediction of fluctuations using SVR. Then, the accuracy of the fluctuation prediction by SVR and by GNM is further examined. Finally, we also consider the structural variation of models by NMR in comparison with prediction as well as the fluctuations observed in MD trajectories.

### Correlation of static structural features of proteins with fluctuations

In Table I, we compared the correlation coefficient of individual structural features with the fluctuation of residues observed in the MD trajectories. Eight different distance cutoff values, 6, 8, to 16 Å, were used for the residue contact number. The top of the table shows the correlation of the *B*-factor (0.484). Interestingly, several static structural features, namely, the distance to the center of mass and the contact number computed with the cutoff of 12–22 Å, have more significant correlation with the fluctuations than the *B*-factor. Among the static features, the largest correlation coefficients were observed for the residue contact number (15 and 16 Å). These results indicate that the motion of chains in the MD trajectories is better captured by the coarse-grained topological structures of proteins rather than the *B*-factor.

As a reference, we also show the correlation of the fluctuations predicted by GNM (bottom rows of Table I). GNM showed higher correlation than the other structural features. Note that GNM actually simulates dynamic motion of protein structures; thus, it has a different nature from the other static features compared in the table. Consistently, with the previous work by Yang *et al.*,[38] GNM without using a distance cutoff showed higher correlation than GNM with a distance cutoff.

Because the residue contact number (with a 16 Å cutoff) and the square of distance to the center of mass showed two largest correlation coefficients among the static structure features examined, we used these two features as the basis for combinations of input features for training SVR in the next section.

**Table I**
Correlation Coefficients Between Structural Features and Fluctuations

| Structural features | Number of proteins with P-*value* < *0.05 (%)*[a] | Avg. corr. coeff.[b] |
|---|---|---|
| B-factor | 565 (95.4) | 0.484 (0.504) |
| Distance to center of mass | 584 (98.6) | 0.509 (0.514) |
| Square of distance to center of mass | 586 (99.0) | 0.545 (0.549) |
| Contact number (cutoff 6 Å) | 571 (96.5) | −0.374 (−0.384) |
| Contact number (8 Å) | 591 (99.8) | −0.480 (−0.481) |
| Contact number (12 Å) | 590 (99.7) | −0.554 (−0.556) |
| Contact number (15 Å) | 587 (99.2) | **−0.568 (−0.571)** |
| Contact number (16 Å) | 571 (96.5) | **−0.567 (−0.571)** |
| Contact number (18 Å) | 587 (99.2) | −0.562 (−0.565) |
| Contact number (20 Å) | 585 (98.8) | −0.555 (−0.559) |
| Contact number (22 Å) | 584 (98.6) | −0.545 (−0.551) |
| Accessible Surface Area (ASA)[c] | 580 (98.0) | 0.404 (0.407) |
| ASA normalized | 590 (99.7) | 0.476 (0.477) |
| Residue depth (residue mean)[d] | 559 (94.4) | −0.352 (−0.371) |
| Residue depth (Cα) | 553 (93.4) | −0.339 (−0.359) |
| Half upper sphere exposure (Cα)[e] | 568 (95.9) | −0.385 (−0.398) |
| Half lower sphere exposure (Cα) | 567 (95.8) | −0.389 (−0.402) |
| Half upper sphere exposure (Cβ) | 537 (90.7) | −0.339 (−0.363) |
| Half lower sphere exposure (Cβ) | 561 (94.8) | −0.383 (−0.399) |
| Prediction by GNM (cutoff 16 Å)[f] | 586 (99.0) | 0.643 (0.648) |
| Prediction by GNM (no cutoff) | 591 (99.8) | 0.646 (0.646) |

The largest correlation coefficients among the static structural features are highlighted in bold.
[a]The number of proteins that have significant correlation coefficient to the fluctuations (with P-value < 0.05) are counted. The total number of trajectories (proteins) is 592.
[b]The average value calculated only for the subset of proteins with P-value < 0.05 is shown in the parentheses.
[c]Accessible surface area (Å$^2$) of amino acid residues without normalization. The next row is the correlation with the normalized accessible surface area.
[d]The residue depth computed as the average distance for each atom in the residue and the distance for the Cα atom (next row).
[e]The lower/upper half-sphere exposure of a residue using the Cα or the Cβ atom to determine the position of the plane which cut the sphere to half.
[f]Fluctuations predicted by GNM [Eq. (6)].

### SVR models for predicting residue fluctuation using static structure features

Next, we used SVR to predict the fluctuation of residue positions in the MD trajectories using various combinations of static structural features. Fluctuation predictions by GNM (at the bottom of Table I) were not included as features. Fivefold cross validation was performed, in which SVR parameters were trained on four-fifths of the dataset, while prediction was made for the rest of the one-fifth of the dataset. This procedure was repeated five times to make prediction for all data in the dataset. Starting from the combination of the residue contact number (with 16 Å cutoff) and the square of distance to the center of mass, which are the two features that showed the highest correlation with fluctuations (Table I), 17 different feature combinations were tested by adding one feature at a time (Table II).

Among the 17 feature combinations examined, all except for two (the feature set 1 and set 17) showed higher correlation with actual fluctuations than GNM (Table I). The largest correlation coefficient, 0.669, was achieved for the feature set 15, which uses the residue contact numbers with different distance cutoffs. In terms of average RMS, all the feature combinations predicted residue fluctuations within an RMS of 1.1 Å, ranging from 1.042 to 1.092 Å. The smallest RMS was achieved for feature sets 6, 7, 12, 13, and 14, which combine the residue contact numbers, the square distance from the center of mass, and the *B*-factor. Sets 6 and 7

**Table II**
Summary of Fluctuation Prediction Using SVR Models with Different Feature Combinations

| Feature set | Features used[a] | Number of proteins with P-*value* < *0.05 (%)* | Average corr. coeff.[b] | RMS (Å)[c] |
|---|---|---|---|---|
| 1 | C(16), $D^2$ | 584 (98.6) | 0.638 (0.644) | 1.075 |
| 2 | C(16), $D^2$, B | 587 (99.2) | 0.654 (0.658) | 1.067 |
| 3 | C(16), $D^2$, B, C(18) | 587 (99.2) | 0.655 (0.659) | 1.060 |
| 4 | C(16), $D^2$, B, C(18), Sec | 589 (99.5) | 0.661 (0.664) | 1.048 |
| 5 | C(16), $D^2$, B, C(18), Res-type | 586 (99.0) | 0.652 (0.657) | 1.063 |
| 6 | C(16), $D^2$, B, C(18), Sec, C(12) | 589 (99.5) | 0.665 (0.668) | **1.042** |
| 7 | C(16), $D^2$, B, C(18), Sec, C(12), C(8) | 588 (99.3) | 0.667 (0.668) | **1.042** |
| 8 | C(16), $D^2$, C(18), C(12), C(8), C(6) | 588 (99.3) | 0.656 (0.660) | 1.053 |
| 9 | C(16), $D^2$, B, C(18), C(12), C(8), C(6) | 588 (99.3) | 0.666 (0.669) | 1.045 |
| 10 | C(16), $D^2$, B, C(18), C(12), C(8), C(6), Sec | 589 (99.5) | 0.665 (0.667) | 1.043 |
| 11 | C(16), $D^2$, B, C(18), C(12), C(8), C(6), Acc | 587 (99.2) | 0.665 (0.669) | 1.045 |
| 12 | C(16), $D^2$, B, C(18), C(12), C(8), C(6), C(20) | 588 (99.3) | 0.666 (0.670) | **1.042** |
| 13 | C(16), $D^2$, B, C(18), C(12), C(8), C(6), C(20), C(22) | 588 (99.3) | 0.667 (0.670) | **1.042** |
| 14 | C(16), $D^2$, B, C(18), C(12), C(8), C(6), C(15), C(20), C(22) | 588 (99.3) | 0.666 (0.670) | **1.042** |
| 15 | C(16), B, C(18), C(12), C(8), C(6), C(20), C(22) | 588 (99.3) | **0.669 (0.673)** | 1.073 |
| 16 | C(16), C(18), C(12), C(8), C(6), C(15), C(20), C(22) | 587 (99.2) | 0.660 (0.665) | 1.092 |
| 17 | C(16), B, C(18), C(12), C(8), C(6), C(20), C(22), HP | 587 (99.2) | 0.647 (0.651) | 1.092 |

The largest correlation coefficients among the static structural features are highlighted in bold.
[a]C(x), the residue contact number with x Å distance cutoff; B, B-factor; $D^2$, square of the distance between the Cα atom to the protein center of mass; Sec, the secondary structure; Acc, normalized accessible surface area; HP, the number of hydrophilic/hydrophobic contacts, Res-Type, amino acid type of residues.
[b]The average correlation coefficients between predicted and actual fluctuations. Values calculated only for the subset of proteins that have significant correlation with P-value < 0.05 is shown in the parentheses.
[c]The RMS [Eq. (7)] was averaged over all the proteins in the dataset.

additionally used information about the secondary structure. The RMS and the average correlation coefficients (Table II) correlate moderately with a correlation coefficient of 0.627 (Fig. 3). Figure 4 shows the distribution of the average correlation coefficients between predicted and actual fluctuations [Fig. 4(A)] and the average RMS [Fig. 4(B)] for each protein, which were predicted using feature set 12. Remarkably, the majority (70%) of proteins fluctuations were predicted within an RMS of 1.0 Å. The strong advantage of the developed SVR models is that

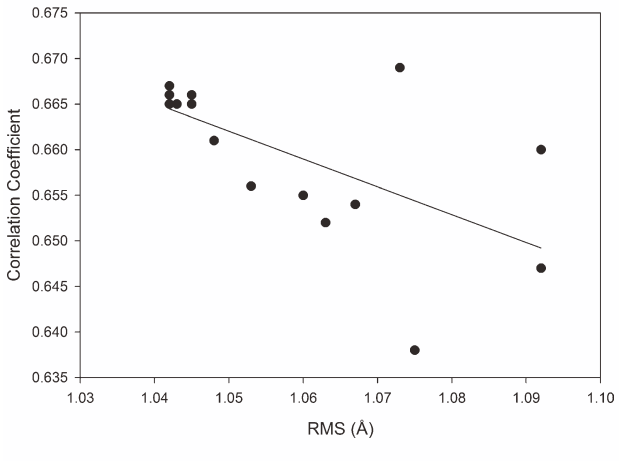



**Figure 3**
The average correlation coefficient and RMS of predicted and actual fluctuations. Predictions were made with SVR using 17 different feature combinations (Table II).

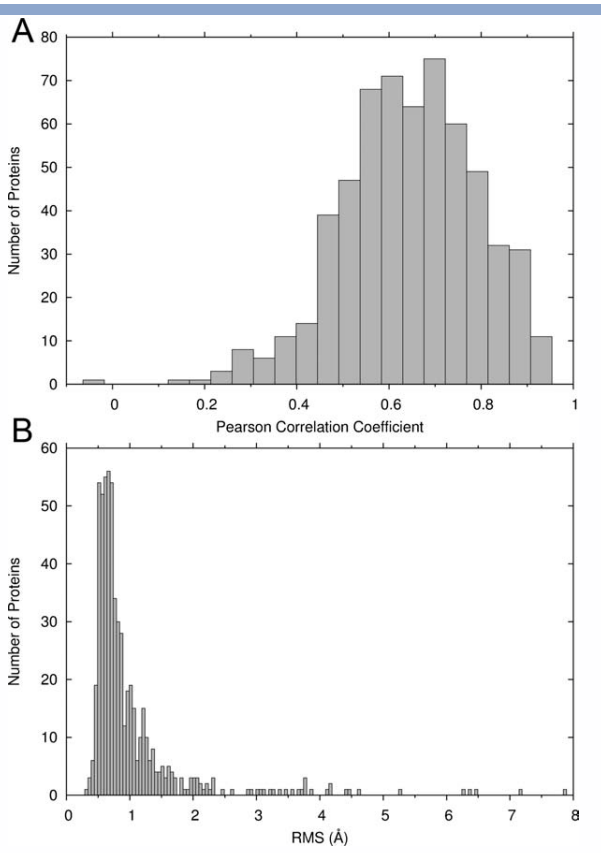



**Figure 4**
Distribution of (**A**), corelation coefficients; (**B**), RMS (Å) of predicted and actual fluctuations computed for 592 proteins in the dataset.

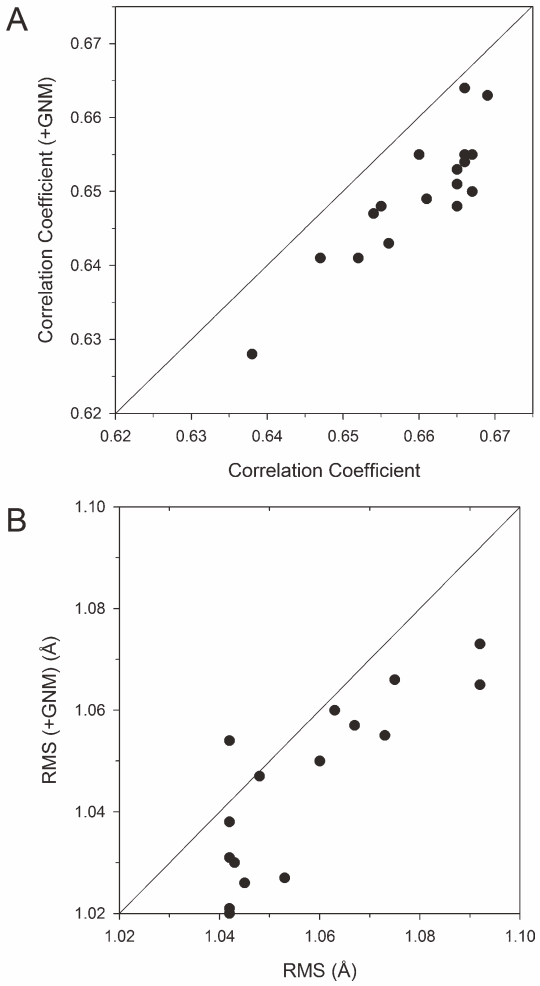



**Figure 5**

Comparison of the prediction performance with and without using GNM as a feature. $\langle(\Delta R_i)^2\rangle$ predicted by GNM was added to each SVR feature set listed in Table II. (**A**) Average correlation coefficient; (**B**) average RMS predicted by SVR with and without $\langle(\Delta R_i)^2\rangle$ from GNM are plotted.

they predict the real value of fluctuation, unlike GNM, which predicts only the relative magnitude of residue fluctuations that need to be rescaled to obtain actual fluctuation values.

## Incorporating dynamic features to SVR models

We further investigated whether adding GNM as an input feature can improve fluctuations prediction with SVR. We used $\langle(\Delta R_i)^2\rangle$ for the fluctuations from GNM [Eq. (6)] without a distance cutoff, because it has higher correlation with the actual fluctuations than $\sqrt{\langle(\Delta R_i)^2\rangle}$ does. To each of the feature sets examined in Table II, we added $\langle(\Delta R_i)^2\rangle$ predicted by GNM and performed five-fold cross validation. The resulting fluctuation prediction with and without GNM was compared in terms of the correlation coefficient [Fig. 5(A)] and the RMS [Fig. 5(B)] with the actual fluctuations.

Adding GNM in the feature set made slight improvement in the RMS of the predicted fluctuations [Fig. 5(B)] except for one case (feature set 12), lowering RMS on average by 0.010. However, small consistent deterioration of the correlation coefficient was observed [Fig. 5(A)] when GNM was added. The average decrease in the correlation coefficient is 0.013. Thus, GNM did not make significant contribution to improving fluctuation prediction.

## Comparison of SVR model prediction results with *B*-factor fluctuation values

In Figure 6, we show four examples of actual and predicted fluctuations as well as fluctuations derived from the *B*-factors. For residue $i$ with a *B*-factor of $B_i$, the fluctuation is defined as

$$\sqrt{\langle(\Delta R_i)^2\rangle}^{\text{Bfactor}} = \sqrt{\frac{3B_i}{8\pi^2}}. \qquad (8)$$

The fluctuations from the *B*-factor were also rescaled to achieve a smaller RMS with the actual fluctuations (i.e., fluctuations from MD trajectories) as follows

$$\sqrt{\langle(\Delta R_i)^2\rangle}_{\text{rescaled}} = \sqrt{\langle(\Delta R)^2\rangle}_{\text{min}} + \alpha\left(\sqrt{\langle(\Delta R)^2\rangle}_{\text{max}} - \sqrt{\langle(\Delta R)^2\rangle}_{\text{min}}\right)\frac{\sqrt{\langle(\Delta R_i)^2\rangle}^{\text{Bfactor}} - \sqrt{\langle(\Delta R)^2\rangle}_{\text{min}}^{\text{Bfactor}}}{\sqrt{\langle(\Delta R)^2\rangle}_{\text{max}}^{\text{Bfactor}} - \sqrt{\langle(\Delta R)^2\rangle}_{\text{min}}^{\text{Bfactor}}}, \qquad (9)$$

where $\sqrt{\langle(\Delta R)^2\rangle}_{\text{max}}$ and $\sqrt{\langle(\Delta R)^2\rangle}_{\text{min}}$ are the maximum and the minimum values of actual fluctuations, and $\sqrt{\langle(\Delta R)^2\rangle}_{\text{max}}^{\text{Bfactor}}$ and $\sqrt{\langle(\Delta R)^2\rangle}_{\text{min}}^{\text{Bfactor}}$ are the maximum and the minimum fluctuation values computed from *B*-factor values [Eq. (8)] in the protein. $\alpha$ is a weighting factor explored from 0.1 to 1.0 with an interval of 0.1 to seek smaller RMS for the actual fluctuations (Table III). In Figure 6, $\alpha$ is set to 1.0 for the plots of "Fluctuation from *B*-factor, rescaled." Note that this rescaling obviously changes the RMS but does not change the correlation coefficient to the actual fluctuation. The acutal fluctuations in the MD trajectories are defined by Eq. (2), and predictions were made using feature set 15 in Table II. The right panel of

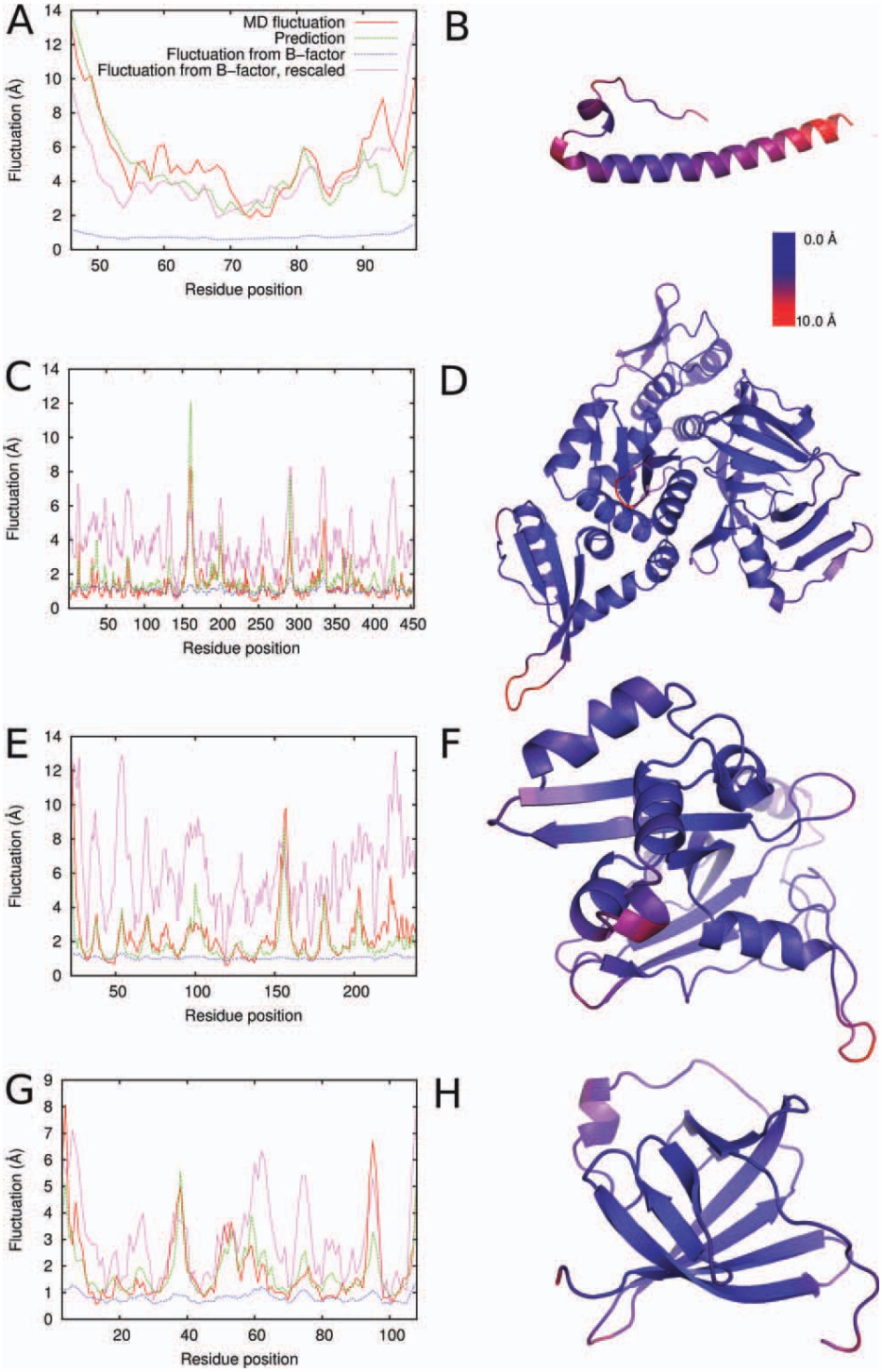



**Figure 6**

Examples of predicted fluctuations in comparison with *B*-factor-derived fluctuations and MD simulation fluctuations. Left panels show the values of fluctuations: red, fluctuations observed in the MD trajectories; green, predicted fluctuations; dotted blue line, fluctuations computed from *B*-factors; dotted magenta line, rescaled fluctuations from *B*-factors ($\alpha = 1.0$). The correlation coefficients and RMS are summarized in Table III. Right-hand panels show the magnitude of fluctuations in a color scale with blue indicating lower fluctuations and red for higher fluctuations. **A**, **B**, 1mof; **C**, **D**, 1dq3; **E**, **F**, 1gpc; **G**, **H**, 1a1x.

**Table III**
Correlation Coefficients and RMS of the Four Example Predictions

| | Correlation coefficient | | RMS (Å) | | | |
|---|---|---|---|---|---|---|
| PDB ID | B-factor | Prediction | B-factor | B-factor, rescaled $\alpha = 1.0$[a] | B-factor, rescaled ($\alpha$)[b] | Prediction |
| 1mof | 0.69 | 0.80 | 4.92 | 1.91 | 1.91 (1.0) | 1.55 |
| 1dq3 | 0.50 | 0.81 | 0.94 | 2.64 | 0.85 (0.4) | 0.71 |
| 1gpc | 0.55 | 0.78 | 1.93 | 4.32 | 1.42 (0.4) | 1.04 |
| 1alx | 0.61 | 0.82 | 1.60 | 1.72 | 1.09 (0.6) | 0.79 |

The data correspond to plots at the left panels in Figure 6.
[a]Fluctuations computed from B-factor were rescaled with $\alpha = 1.0$ in [Eq. (9)]. This value corresponds to the curve "Fluctuation from B-factor, rescaled" in Figure 6.
[b]Fluctuations computed from B-factor were rescaled with the weight factor $\alpha$ [Eq. (9)] ranging from 0.1 to 1.0 with an interval of 0.1. Then the smallest RMS obtained is shown together with the used $\alpha$ value in the parentheses.

each protein visualizes the magnitude of actual fluctuations in a color scale from blue to red with blue showing small while red for large fluctuation.

The first example, retrovirus coat protein (PDB ID: 1mof) [Fig. 6(A,B)], exhibits a large fluctuation at two termini and at the end of the long helix. Prediction by SVR captured fluctuating residues and the magnitude fairly well with a correlation coefficient of 0.80 and an RMS of 1.55 Å. The fluctuations derived from *B*-factor have lower correlation with the actual fluctuations (correlation coefficient of 0.69) with a larger RMS of 1.91 Å even after rescaling. In the second example [Fig. 6(C,D)] of homing endonuclease PI-PfuI (PDB ID: 1dq3), overall fluctuation is not large but shows high peaks of fluctuation at loop regions. The predicted fluctuations have a correlation coefficient of 0.81 while the fluctuations from *B*-factor have a moderate correlation of 0.50. The third example, DNA-binding protein gp32 (PDB ID: 1gpc) [Fig. 6(E,F)], has the largest fluctuation at the loop of residues 150–160 and over 3 Å fluctuation at the other loop regions, which are captured well by the prediction. Predicted fluctuations have a correlation coefficient of 0.78 and a small RMS of 1.04 Å. In contrast, the correlation of fluctuations from *B*-factor is 0.55 with a larger RMS of 1.93 Å. The last example, MTCP-1 (PDB ID: 1a1x) [Fig. 6(G,H)], is a β-barrel protein with a long loop at residues 50–60. Relatively large fluctuation was observed at the N-terminus and at the loop regions that connect β-strands (e.g., residues 35–40), which are well predicted. The overall RMS of the prediction is 0.79 Å, and the correlation coefficient with the actual fluctuations is 0.82, better than the fluctuations derived from *B*-factors.

Consistent with Table I, the fluctuations from *B*-factors correlate only moderately with the actual fluctuations. Fluctuations computed from *B*-factors using Eq. (8) have always a larger RMS than the SVR prediction. The agreement of the fluctuations from *B*-factors can be improved if it is rescaled individually for each protein as shown in the second column from the right in Table III; however, the value of the optimal scaling factor $\alpha$ differs from protein to protein and thus cannot be known beforehand. In contrast, our prediction by SVR has a significantly higher correlation with the actual prediction, and it predicts the real value of the fluctuations satisfactorily without any rescaling.

**Table IV**
Comparison of Fluctuations of NMR Models, MD, and Our Prediction

| Compared data | Number of proteins with P-value $< 0.05$ (%) | Corr. coeff. | RMS (Å) |
|---|---|---|---|
| NMR versus MD | 136 (97.1) | 0.651 (0.667) | 2.425 |
| NMR versus prediction | 138 (98.6) | 0.739 (0.747) | 1.808 |
| MD versus prediction | 138 (98.6) | 0.686 (0.693) | 2.165 |

Hundred and forty nonredundant proteins in the MoDEL database were used whose structures were determined by NMR.

### MD fluctuations and fluctuations from NMR models

The MoDEL database also contains simulations of protein structures determined by NMR. We selected 140 nonredundant protein structures determined by NMR that contain more than 10 models in their PDB files. Redundant proteins were removed by considering sequence identity according to the PISCES database.[63] Using the 140 proteins, we compared fluctuations observed in the NMR models, MD trajectories, and the predicted fluctuations. The results are summarized in Table IV. The fluctuation prediction was carried out using feature set 16, which does not contain the *B*-factor term (NMR structures do not have *B*-factors).

It is shown that the prediction has a significant correlation (0.739) with the structural variation of the models derived from NMR. Interestingly, the correlation coefficient between the prediction and NMR is highest among the other two pairs, prediction versus MD and NMR versus MD.

## CONCLUSION

We used a large number of MD trajectories of nonhomologous proteins as references and examined static structural features of the proteins that are most relevant

to fluctuations. We examined the correlation of individual structural features with fluctuations and then investigated effective combinations of features for SVR to predict the real value of fluctuation of residues. The main findings of this work are summarized as follows. First of all, two types of structural features, the distance to the center of mass of the protein and the residue contact number, showed a higher correlation coefficient with fluctuations than *B*-factor does. Combinations of static features used as input for SVR achieved accurate prediction of fluctuations with a correlation coefficient of 0.67 and RMS of 1.042 Å. This correlation coefficient is higher than GNM to the actual fluctuation. Our method predicts the structural variation of NMR models also well. The current study demonstrates that flexibility of proteins is inherently coded in coarse-grained static protein structural features, even more than in the crystallographic *B*-factors. Thus, protein motion is determined by its static structure that is coded by its sequence, which could be considered as an extension of the Anfinsen's dogma.[75] Indeed, series of studies on GNM has also demonstrated that motion of a protein is determined by its structure. However, the current work further shows that static structural features can predict the real value of fluctuations, which GNM has not been shown to be able to do. As the importance of protein dynamics has been more recognized for biological function, the prediction method we developed has also a practical value in the wide areas of biology and biotechnology.

## ACKNOWLEDGMENTS

The authors thank Jordi Camps (Centre Nacional d'Anàlisis Genòmica, Spain) and Tim Meyer (Institute for Research in Biomedicine, Spain) for help with the PCAsuite software and the MoDEL database.

## REFERENCES

1. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. Science 2006;311:347–351.
2. Todd AE, Marsden RL, Thornton JM, Orengo CA. Progress of structural genomics initiatives: an analysis of solved target structures. J Mol Biol 2005;348:1235–1260.
3. Westbrook J, Feng Z, Chen L, Yang H, Berman HM. The Protein Data Bank and structural genomics. Nucleic Acids Res 2003;31:489–491.
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
5. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2006;34:D291–D295.
6. Kihara D, Skolnick J. Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. Proteins 2004;55:464–473.
7. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol 2008;18:342–348.
8. Chen H, Kihara D. Effect of using suboptimal alignments in template-based protein structure prediction. Proteins 2011;79:315–334.
9. Das R, Baker D. Macromolecular modeling with rosetta. Annu Rev Biochem 2008;77:363–382.
10. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
11. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA 2001;98:10125–10130.
12. Kihara D, Zhang Y, Lu H, Kolinski A, Skolnick J. Ab initio protein structure prediction on a genomic scale: application to the *Mycoplasma genitalium* genome. Proc Natl Acad Sci USA 2002;99:5993–5998.
13. Borreguero JM, Skolnick J. Benchmarking of TASSER in the ab initio limit. Proteins 2007;68:48–56.
14. Trojanowski S, Rutkowska A, Kolinski A. TRACER: a new approach to comparative modeling that combines threading with free-space conformational sampling. Acta Biochim Polym 2010;57:125–133.
15. Venkatraman V, Sael L, Kihara D. Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. Cell Biochem Biophys 2009;54:23–32.
16. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein-RNA interactions. J Struct Biol, DOI: 10.1016/j.jsb.2011.10.001.
17. Ritchie DW. Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci 2008;9:1–15.
18. Hillisch A, Pineda LF, Hilgenfeld R. Utility of homology models in the drug discovery process. Drug Discov Today 2004;9:659–669.
19. Takeda-Shitaka M, Takaya D, Chiba C, Tanaka H, Umeyama H. Protein structure prediction in structure based drug design. Curr Med Chem 2004;11:551–558.
20. Teilum K, Olsen JG, Kragelund BB. Functional aspects of protein flexibility. Cell Mol Life Sci 2009;66:2231–2247.
21. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. J Mol Graph Model 2001;19:26–59.
22. Hammes GG, Benkovic SJ, Hammes-Schiffer S. Flexibility, diversity, and cooperativity: pillars of enzyme catalysis. Biochemistry 2011;50:10422–10430.
23. Chiti F, Dobson CM. Amyloid formation by globular proteins under native conditions. Nat Chem Biol 2009;5:15–22.
24. Zacharias M. Accounting for conformational changes during protein-protein docking. Curr Opin Struct Biol 2010;20:180–186.
25. Mandell DJ, Kortemme T. Backbone flexibility in computational protein design. Curr Opin Biotechnol 2009;20:420–428.
26. Lassila JK. Conformational diversity and computational enzyme design. Curr Opin Chem Biol 2010;14:676–682.
27. Lill MA. Efficient incorporation of protein flexibility and dynamics into molecular docking simulations. Biochemistry 2011;50:6157–6169.
28. Debye P. Interferenz von Röntgenstrahlen und Wärmebewegung. Annal Phys 1913;348:49–92.
29. Eastman P, Pellegrini M, Doniach S. Protein flexibility in solution and in crystals. J Chem Phys 1999;110:10141–10152.
30. Ishima R, Torchia DA. Protein dynamics from NMR. Nat Struct Biol 2000;7:740–743.
31. Baldwin AJ, Kay LE. NMR spectroscopy brings invisible protein states into focus. Nat Chem Biol 2009;5:808–814.
32. Nilges M, Habeck M, O'Donoghue SI, Rieping W. Error distribution derived NOE distance restraints. Proteins 2006;64:652–664.
33. Chalaoux FR, O'Donoghue SI, Nilges M. Molecular dynamics and accuracy of NMR structures: effects of error bounds and data removal. Proteins 1999;34:453–463.
34. Brooks B, Karplus M. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 1983;80:6571–6575.

35. Haliloglu T, Bahar I, Erman B. Gaussian dynamics of folded proteins. Phys Rev Lett 1997;79:3090–3093.
36. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 1996;77:1905–1908.
37. Bahar I, Erman B, Haliloglu T, Jernigan RL. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. Biochemistry 1997;36:13512–13523.
38. Yang L, Song G, Jernigan RL. Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci USA 2009;106: 12347–12352.
39. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J Comp Chem 1997;18:849–873.
40. Kolinski A. Protein modeling and structure prediction with a reduced representation. Acta Biochim Polym 2004;51:349–371.
41. He Y, Liwo A, Weinstein H, Scheraga HA. PDZ binding to the BAR domain of PICK1 is elucidated by coarse-grained molecular dynamics. J Mol Biol 2011;405:298–314.
42. Kmiecik S, Kolinski A. Characterization of protein-folding pathways by reduced-space modeling. Proc Natl Acad Sci USA 2007;104: 12330–12335.
43. Kmiecik S, Kolinski A. Folding pathway of the b1 domain of protein G explored by multiscale modeling. Biophys J 2008;94:726–736.
44. Kmiecik S, Kolinski A. Simulation of chaperonin effect on protein folding: a shift from nucleation-condensation to framework mechanism. J Am Chem Soc 2011;133:10283–10289.
45. Kondrashov DA, Cui Q, Phillips GN Jr. Optimization and evaluation of a coarse-grained model of protein motion using X-ray crystal data. Biophys J 2006;91:2760–2767.
46. Lin TL, Song G. Generalized spring tensor models for protein fluctuation dynamics and conformation changes. BMC Struct Biol 2010;10 (Suppl 1):S3.
47. Micheletti C, Carloni P, Maritan A. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. Proteins 2004;55:635–645.
48. Canino LS, Shen T, McCammon JA. Changes in flexibility upon binding: application of the self-consistent pair contact probability method to protein-protein interactions. J Chem Phys 2002;117: 9927–9933.
49. Pandey BP, Zhang C, Yuan X, Zi J, Zhou Y. Protein flexibility prediction by an all-atom mean-field statistical theory. Protein Sci 2005;14:1772–1777.
50. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. On the relation between residue flexibility and local solvent accessibility in proteins. Proteins 2009;76:617–636.
51. Shih CH, Huang SW, Yen SC, Lai YL, Yu SH, Hwang JK. A simple way to compute protein dynamics without a mechanical model. Proteins 2007;68:34–38.
52. Kloczkowski A, Jernigan RL, Wu Z, Song G, Yang L, Kolinski A, Pokarowski P. Distance matrix-based approach to protein structure prediction. J Struct Funct Genom 2009;10:67–81.
53. Bornot A, Etchebest C, De Brevern AG. Predicting protein flexibility through the prediction of local structures. Proteins 2011;79: 839–852.
54. Gu J, Gribskov M, Bourne PE. Wiggle-predicting functionally flexible regions from primary sequence. PLoS Comput Biol 2006;2:e90.
55. Chen P, Wang B, Wong HS, Huang DS. Prediction of protein B-factors using multi-class bounded SVM. Protein Pept Lett 2007; 14:185–190.
56. Hirose S, Yokota K, Kuroda Y, Wako H, Endo S, Kanai S, Noguchi T. Prediction of protein motions from amino acid sequence and its application to protein-protein interaction. BMC Struct Biol 2010;10:20.
57. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins 2005;61:115–126.
58. Meyer T, D'Abramo M, Hospital A, Rueda M, Ferrer-Costa C, Perez A, Carrillo O, Camps J, Fenollosa C, Repchevsky D, Gelpi JL, Orozco M. MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. Structure 2010;18:1399–1409.
59. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem 2005;26:1668–1688.
60. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: algorithms for highly efficient, load-aalanced, and scalable molecular simulation. J Chem Theory Comput 2008;4:435–447.
61. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. J Comput Chem 2005;26:1781–1802.
62. Meyer T, Ferrer-Costa C, Perez A, Rueda M, Bidon-Chanal A, Luque FJ, Laughton CA, Orozco M. Essential dynamics: a tool for efficient trajectory compression and management. J Chem Theory Comput 2006;2:251–258.
63. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
64. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
65. Lin CP, Huang SW, Lai YL, Yen SC, Shih CH, Lu CH, Huang CC, Hwang JK. Deriving protein dynamical properties from weighted protein contact number. Proteins 2008;72:929–935.
66. Halle B. Flexibility and packing in proteins. Proc Natl Acad Sci USA 2002;99:1274–1279.
67. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
68. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.
69. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:641–656.
70. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure 1999;7: 723–732.
71. Sanner M, Olson AJ, Spehner JC. Fast and robust computation of molecular surfaces. Proceedings of 11th ACM Symposium on Computational Geometry, Vancouver, BC, Canada; 1995. ppC6–C7.
72. Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins 2005;59:38–48.
73. Kundu S, Melton JS, Sorensen DC, Phillips GN Jr. Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys J 2002;83:723–732.
74. Chang C-C, Jin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2001;2:27:1–27:27.
75. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.

# E | A CONSISTENT VIEW OF PROTEIN FLUCTUATIONS FROM ALL-ATOM MOLECULAR DYNAMICS AND COARSE-GRAINEDDYNAMICS WITH KNOWLEDGE-BASED FORCE-FIELD

*Jamroz M, Orozco M, Kolinski A, Kmiecik S. (2013)* A Consistent View of Protein Fluctuations from All-atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-based Force-field. *Journal of Chemical Theory and Computation 9(1):119–125.*

JCTC Journal of Chemical Theory and Computation
Article
pubs.acs.org/JCTC

# Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field

Michal Jamroz,[†] Modesto Orozco,[‡,¶] Andrzej Kolinski,[†] and Sebastian Kmiecik*,[†]

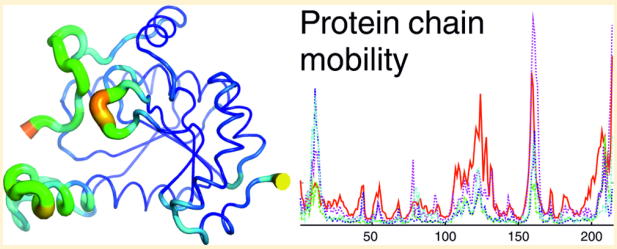
[†]Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

[‡]IRB - BSC Joint Research Program in Computational Biology, Institute for Research in Biomedicine, Josep Samitier 1-5, Barcelona 08028, Spain

[¶]Department of Biochemistry, Universitat of Barcelona, Gran Via de les Corts Catalanes, 585 08007 Barcelona, Spain

Supporting Information

**ABSTRACT:** It is widely recognized that atomistic Molecular Dynamics (MD), a classical simulation method, captures the essential physics of protein dynamics. That idea is supported by a theoretical study showing that various MD force-fields provide a consensus picture of protein fluctuations in aqueous solution [Rueda, M. et al. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 796−801]. However, atomistic MD cannot be applied to most biologically relevant processes due to its limitation to relatively short time scales. Much longer time scales can be accessed by properly designed coarse-grained models. We demonstrate that the aforementioned consensus view of protein dynamics from short (nanosecond) time scale MD simulations is fairly consistent with the dynamics of the coarse-grained protein model - the CABS model. The CABS model employs stochastic dynamics (a Monte Carlo method) and a knowledge-based force-field, which is not biased toward the native structure of a simulated protein. Since CABS-based dynamics allows for the simulation of entire folding (or multiple folding events) in a single run, integration of the CABS approach with all-atom MD promises a convenient (and computationally feasible) means for the long-time multiscale molecular modeling of protein systems with atomistic resolution.




## 1. INTRODUCTION

Protein folding is a very complex process involving very fast local dynamics and long-time scale rearrangements of a large number of atoms. Local fluctuations (side-chains, loops) occur in picoseconds, while global rearrangements (folding/unfolding) require typically milliseconds, even for small globular proteins. No experimental or simulation technique is able to embrace all spatial and temporal scales relevant to process description.[1,2] Thus, complete characterization of the folding process requires proper integration of data from a variety of experimental and computational methods. Recent examples of such integrative characterization involve a description of the smallest systems and time scales[3] as well as large macromolecular machines in motion.[4]

As noted above, folding, and in fact most relevant biological processes involving protein conformational changes, takes place on large time scales (between 10 microseconds and milliseconds or even hours), making most of them inaccessible to atomistic MD simulation. Supercomputer efforts in the past few years established the limit of such simulations to be around 10 microseconds of simulated biological time.[5] Just very recently the 1-millisecond barrier was broken by the Shaw group thanks to a custom-built supercomputer.[6] The 1-ms simulation of folded protein BPTI (58 residues) revealed distinct separation of time-scales: hopping between structurally different conformational states on time scales of the order of 10 microseconds, whereas local relaxations occurred on a time scale at least 1000 times faster. The fast relaxations were found to originate primarily from side chain motions, whereas the slow relaxations corresponding to transitions between well separated basins originated mostly from backbone motions.[6] Shaw's group also succeeded in modeling the folding pathway of a 35-residue protein[6] and later continued folding simulation studies of larger and more complex fast-folding proteins.[7] The atomic MD simulations (over periods ranging between 100 microseconds and 1 ms) of 11 out of the 12 structurally diverse proteins studied (ranging from 10 to 80 residues) resulted in spontaneous and repeated folding to their experimentally determined native structures. Interestingly, for most cases, folding proceeded along a single, dominant route, where additional structural elements were formed in a well-defined sequence.[7] What is important is that these unique simulations (with respect to protein size and simulation time) were performed using a single force-field that was able to consistently fold a substantial number of proteins, representing major

Received: October 3, 2012
Published: December 3, 2012

   dx.doi.org/10.1021/ct300854w | J. Chem. Theory Comput. 2013, 9, 119−125

structural classes, to their native states. This result suggests that current MD force-fields may be accurate enough for conducting long time-scale MD simulations. However, another study of the same group, using different force-fields to folding of the villin headpiece,[8] showed that even all studied force-fields were able to fold the protein with folding rates consistent with the experiment, the observed folding pathways depended on the choice of the force-field and the properties of the unfolded state were substantially different among various force-fields. Importantly, a number of other studies (applying atomistic MD and explicit representation of water molecules) confirmed a possibility to fold a protein into its native tertiary structure[6,9−13] and also the inconsistency of different force-fields in the description of a folding pathway.[6,14,15]

While MD simulations of large structural rearrangements (such as entire folding processes) showed to be force-field dependent, the simulations of near-native dynamics seem to be essentially force-field independent, as shown by Orozco and colleagues.[16] The authors examined the consistency of different force-fields in the description of near-native protein dynamics (by state-of-the-art atomistic MD simulations with explicit water). The analysis revealed that most of the dynamics behavior is force-field independent. The four most popular force-fields were used: AMBER[17,18] (A), CHARMM[19,20] (C), GROMOS[21,22] (G), and OPLS[23,24] (O), in a massive supercomputer project for proteins with different folds. MD trajectories from the different force-fields provided a consensus picture of near-native protein dynamics by classical atomic MD in conditions close to physiological.[16] In this work, we use these trajectories as the reference data for comparison with our simulations conducted by a coarse-grained protein model with stochastic dynamics and statistical potentials − the CABS model. This work is a continuation of our previous studies of testing the capability of the CABS model which are successful examples of protein folding simulations from fully denatured to the near-native state.[25−29]

## 2. MATERIALS AND METHODS

**2.1. Protein Data Set and MD Data.** We used all the currently available MD trajectories from the Rueda et al.[16] MD dynamics analysis deposited in the microMoDEL subset of the MoDEL database.[30] The protein data set is listed in Table S1. For all the proteins in the data set 10-ns simulation MD runs were performed with explicit water (the TIP3P water model was used for A, C, and O simulations, and the SPC water model for G simulations) at constant pressure (1013.25 hPa) and temperature (300 K) using standard coupling schemes (the same in all cases).[16]

Experimental mobility profiles (Figure 2 and Figure S2) were derived from crystallographic B-factors ($\langle R^2\rangle_i = (3B_i)/(8\pi^2)$, where $B$ is the B-factor) or multimodel NMR structures (calculated in the same way as for trajectories, see eq 1). In the cases of NMR solved structures: 1BSN, 1SDF, 1IL6, and 1I6F (deposited in the PDB database as a single model), equivalent multimodel PDB data was used (1BSH, 2SDF, 2IL6, and 1I6G, respectively), except for 1FVQ for which multimodel data were not available.

**2.2. CABS Dynamics.** Coarse-grained models, employing united atom representation, offer substantial extension of the time scales of simulated systems compared to those of all-atom models.[2,31−33] The CABS model (described in detail elsewhere[34]) employs coarse-grained representation of a polypeptide chain that uses up to four atoms per residue. These are $C^\alpha$ and $C^\beta$ atoms and two virtual pseudoatoms: one placed in the center of mass of a side-chain and the other in the center of the $C^\alpha-C^\alpha$ virtual bond (see Figure 1). The CABS force-field is

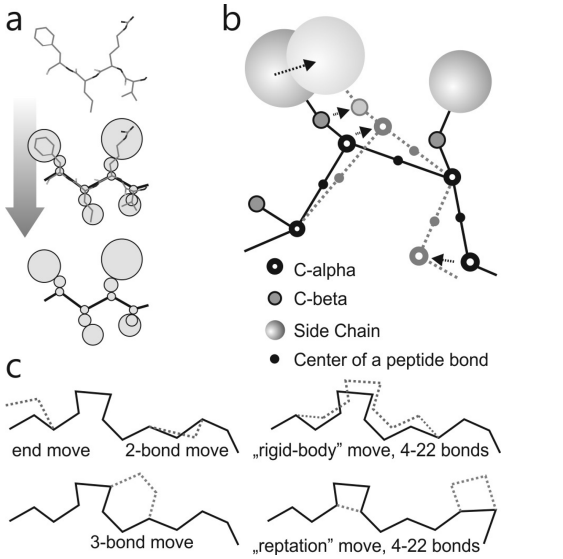



**Figure 1.** CABS model overview: (a) reduced representation, (b) single C-alpha kink move, (c) schematic illustration of larger scale moves of the Monte Carlo scheme.

derived from statistical regularities seen in known protein structures, and it includes side-chain−side-chain mean field potentials, coarse-grained models of main chain hydrogen bonds, and local peptide-chain geometric preferences. The solvent effect is accounted for in an implicit fashion through protein structure statistics used in the derivation of the CABS force-field. The dynamics of CABS-based coarse-grained proteins is simulated by a random series of local conformational transitions (controlled by a Monte Carlo method). Thus, very short-time dynamics, relevant to small-distance local geometric changes, is not physical. However, long series of such random local transitions (modulated by the model force-field) define realistic long-time dynamics, as shown in previous reports.[25−29] Apart from the application to protein dynamics, it is worth noting that the CABS model proved very efficient and accurate in numerous protein structure prediction applications: *de novo* or comparative modeling (e.g.: ranked the best, or one of the best, among other approaches in CASP6 blind prediction experiments[35]) or loop modeling.[36] Importantly, the spatial resolution of CABS predictions enables computationally inexpensive conversion to realistic all-atom models (as shown in the application to high-resolution structure prediction[26] or all-atom description of a folding pathway[27]).

**2.3. CABS Simulation Setup.** For the whole protein data set, more than a hundred simulation setups were performed to test various temperature values, scaling coefficients of force-field components, and versions of distance restraints (unrestrained simulations were also performed) to find the best correlation coefficient for residue fluctuation profiles with the MD trajectories. The highest Spearman's correlation coefficient was found for the simulations with local native-like restraints put exclusively on pairs of residues under two conditions: (1) the distance between their $C^\alpha$ atoms was smaller than 8 Å, (2) both residues were assigned to belong to secondary structure

Journal of Chemical Theory and Computation Article

elements. Therefore, loop regions were completely unrestrained and regions of secondary structure locally only. The applied distance restraints softly penalized the position of restrained residues if their distance differed from the distance in the native structure by more than 1 Å.

**2.4. Analysis of MD and CABS Trajectories.** MD and coarse-grained trajectories were analyzed on the level of $C^\alpha$ trace representation to obtain their structural and dynamics characteristics together with their consistency measures.

The mobility of residue $i$ was defined as

$$\left\langle R_i^2 \right\rangle = \frac{1}{N}\sum_j^N ((p_{j,x}^i - c_{j,x}^i)^2 + (p_{j,y}^i - c_{j,y}^i)^2 + (p_{j,z}^i - c_{j,z}^i)^2) \quad (1)$$

where $j$ - trajectory frame, $i$ - residue index, $c$ - position of the $C^\alpha$ atom in the average structure, and $N$ - number of trajectory models.

The global similarity of structures generated by different MD force-fields and the CABS model was obtained by computing the RMSD between all of the snapshots collected in the two trajectories and related similarity index $\Omega$

$$\Omega_{AB} = (\alpha_{AA} + \alpha_{BB})/2\alpha_{AB} \quad (2)$$

where

$$\alpha_{AB} = \frac{1}{M_A M_B}\sum_i^{M_A}\sum_j^{M_B}\left(\frac{1}{N}\sum_t^{3N}(x_{i,t} - x_{j,t})^2\right)^{1/2} \quad (3)$$

where $N$ is the number of atoms, $M$ is the number of frames in the compared trajectories, and $x$ is the residue coordinate. The similarity index was computed using the Bioshell package.[37]

The mean-square displacement autocorrelation function $acorr(\tau)$ was defined as

$$acorr(\tau) = \left\langle \left\langle R^2 \right\rangle \right\rangle_\tau = \frac{1}{M}\sum_t^{M-\tau}\left(\frac{1}{N}\sum_i^N (\vec{p}_{i,t} - \vec{p}_{i,t+\tau})^2\right) \quad (4)$$

where $p_{i,t}$ - position of residue $i$ at time $t$, $N$ - number of protein residues, $M$ - number of trajectory frames, and $\tau$ - time frame ($\Delta t$).

Global flexibility was shown by the Lindemann's disorder index[38]

$$\Delta_L = \frac{\sqrt{\frac{1}{N}\sum_i^N \left\langle R_i^2 \right\rangle}}{a'} \quad (5)$$

where $N$ is the number of atoms, $a'$ is the most-probable nonbonded near-neighbor distance, and $\langle R_i^2 \rangle$ is the fluctuation of the residue $i$ (see eq 1).[38] Lindemann's disorder index was calculated using the PCASuite package.[39]

Commonly used deformation space overlap was defined using root-mean-square inner product $\gamma$[40]

$$\gamma_{AB} = \frac{1}{n}\sum_{i,j}^n (\vec{v}_i^A \cdot \vec{v}_j^B)^2 \quad (6)$$

where $A$ and $B$ index the two compared methods, $i$ and $j$ index the eigenvectors (ranked on the basis of their contribution to structural variance), and $n$ is the minimum number of eigenvectors needed to explain 90% of total variance.

Deformation space overlap was defined using root-mean-square inner product "s overlap"[41]

$$s(A, B) = 1 - \frac{d(A, B)}{\sqrt{\mathrm{tr}\, A + \mathrm{tr}\, B}} \quad (7)$$

and

$$d(A, B) = [\sum_i^n (\lambda_i^A + \lambda_j^B) - 2\sum_{i,j}^n \sqrt{\lambda_i^A \lambda_j^B}\,(\vec{v}_i^A \cdot \vec{v}_j^B)^2]^{1/2} \quad (8)$$

where $A$ and $B$ index covariance matrices of the two compared methods, tr is the trace of a matrix, $\lambda$ are index eigenvalues, and $v$ are index eigenvectors.

This measure has several advantages over the commonly used subspace overlap[41] (the overlap between the subspaces of the first $n_A$ and $n_B$ eigenvectors of matrix $A$ and $B$, employed also in the study by Rueda et al.[16]) which depends strongly on $n_A$ and $n_B$ and ignores the eigenvalues. "s overlap" metric takes into account differences between eigenvectors with small and large eigenvalues and treats more correctly degenerate subspaces.

## 3. RESULTS AND DISCUSSION

**3.1. Maximizing MD and CABS Convergence.** We started the CABS simulations of the proteins with the optimization of CABS parameters (simulation time, temperature) to obtain the best possible convergence with the available MD data[16] (see Materials and Methods). As a convergence criterion we used the average Spearman's correlation coefficient ($r_s$) for residue mobility between different MD force-fields and the CABS model. The residue mobility reflects the oscillations of the $C^\alpha$ atom of a residue around its mean position (averaged over the whole trajectory, see eq 1).

The highest mean correlations for the completely unrestrained simulations (average over all simulations) between CABS and A, C, G, and O force-fields were the following: 0.62, 0.61, 0.64, 0.63, respectively (see Table S3 for individual protein values). This level of similarity to all-atom MD were also recently achieved by other prediction methods: Support Vector Regression[42] and Gaussian Network Model[43] (mean correlation coefficients respectively: 0.67 and 0.64, as given in ref 42).

Further examination of the CABS mobility profiles revealed, in comparison to the MD trajectories, sometimes smaller stability of the secondary structure elements, particularly visible at elevated temperatures due to long-distance and very fast motions of more flexible parts of protein structures. Furthermore, relying on this observation, we attempted to increase the CABS and MD convergence by repeating the optimization of CABS parameters (simulation time, temperature) and introduction of various types of distance restraints (derived from native structures) to avoid any long-distance and very fast motions of protein structure (see the SI text for optimization procedure of CABS parameters for simulations with distance restraints and predictive power test).

The optimization results showed that the same parameters setup as trained on the whole protein set was found when the



dx.doi.org/10.1021/ct300854w | J. Chem. Theory Comput. 2013, 9, 119−125

Journal of Chemical Theory and Computation Article

method was optimized on randomly chosen half of the protein set. The predictive strength of the method is slightly lower when evaluated on the remaining half of the protein set, than as tested on the whole set ($r_s$ = 0.70 and 0.74, respectively).

The optimal parameters setup, which yielded overall the highest mean correlations (on the level of 0.74), were obtained with weak native-like restraints applied only locally and between coordinates belonging to the secondary structure elements (alpha or beta) (see the SI text for the parameters details). Therefore, loop-forming residues remained completely unrestrained (for the restraints description see Materials and Methods). That was for the setup with significantly higher temperature than the optimal in unrestrained simulations described above. Thus, in comparison to the unrestrained simulations (optimal with regard to temperature and simulation time), the optimal restrained ones allowed for the following: (1) enhanced mobility of at least loop fragments (higher temperature increases the overall acceptance rate of the moves in the Monte Carlo scheme), (2) additional stabilization of the secondary structure and its motifs, and (3) overall decrease in CABS fluctuation level (see the mean RMSD in Table S3 for unrestrained and restrained CABS simulations).

The average correlation coefficients for residue mobility between different MD force-fields and the CABS-simulations with the optimal setup found are listed in Table 1(detailed

**Table 1. Average Spearman's Correlation Coefficient and Mean RMSD (in Brackets) between MDs (A, C, G, O) and CABS Mobility Profiles**[a]

| | A | C | G | O |
|---|---|---|---|---|
| **CABS** | **0.74 (3.12)** | **0.74 (2.84)** | **0.72 (2.91)** | **0.75 (2.92)** |
| A | 1 | 0.80 (1.75) | 0.78 (2.49) | 0.82 (1.76) |
| C | | 1 | 0.75 (2.23) | 0.81 (1.59) |
| G | | | 1 | 0.75 (2.43) |

[a]The mean values for the whole test set are shown. Individual values for each protein are reported in Table S3.

results for each protein are listed in Table S3). Note that, in this manuscript we report the average statistics for the entire protein set (for the most comprehensive comparison of the methods), however the average from the predictive power test (0.7) should be considered as the estimate of the CABS ability to predict fluctuations from the MD (see the SI text for the optimization details). As highlighted above, the mean correlation between CABS and MD force-fields is on the level of 0.7, which is on a slightly lower level with respect to correlations among different MD force-fields (which varied between 0.75 and 0.82). The average similarity between the mobility profiles measured by RMSD (Table 1) shows more significant differences between CABS and MD force-fields (in the range of 2.8−3.1 Å) than among different MD force-fields (1.8−2.5 Å) which is due to higher average residue oscillations observed in CABS than in MD simulations. For the examples of the mobility profiles with the highest (1FAS, $r_s$ = 0.86) and the lowest (1PDO, $r_s$ = 0.49) correlation coefficients between CABS and MD, see Figure 2. For the mobility profiles visualized on example 3D structures, see Figure 3. As analyzed by Rueda et al.,[16] there is a good agreement between X-ray B-factors and MD-derived mobility profiles, which is also the case of the similarity between experimental (X-ray or NMR) and CABS derived fluctuations (see Figure 2 and Figure S2).

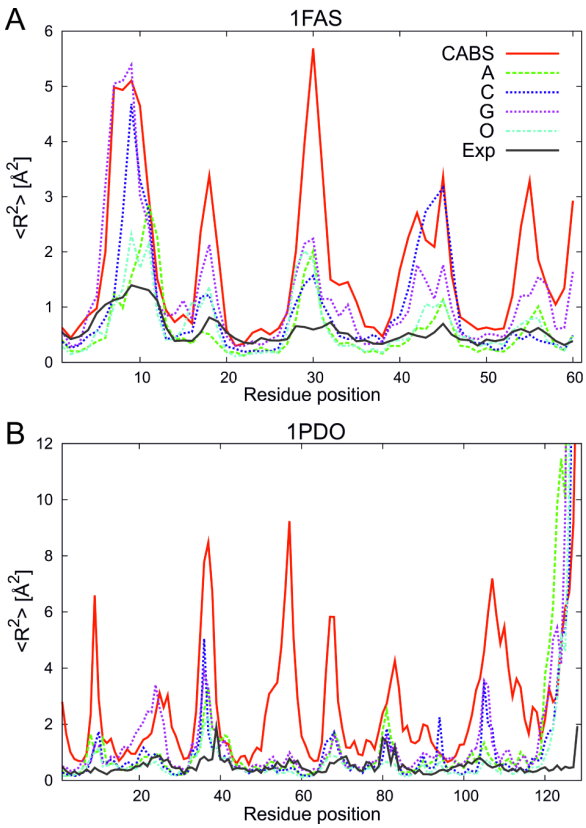



**Figure 2.** Example mobility profiles by the CABS model, different MD force-fields (A, C, G, O) and experimental data (derived from crystallographic B-factors). The profiles are shown for the following: (A) 1FAS (example of the highest correlation between CABS and MD: 0.86) and (B) 1PDO (example of the lowest correlation between CABS and MD: 0.49). See also 1FAS and 1PDO profiles visualized on 3D structures in Figure 3. The profiles for the remaining proteins in the test set are shown in the SI text (Figure S2) together with Spearman's correlation coefficient values for the whole test set (Table S3).

Recently, two similar studies of the suitability of coarse-grained techniques for the prediction of protein dynamics were conducted by Emperador et al.[44,45] The studies tested two Gō-like models:[45] Brownian dynamics (BD[46]) and discrete molecular dynamics (DMD[47]) with a Gō-like Hamiltonian and a DMD model based on a simple pseudophysical force-field[44] (a hybrid between the physical potential and empirical Gō-like model). The simulation results were compared to fully atomistic MD simulations (the same as used in our study). The comparison showed that the coarse-grained models delivered in general similar protein dynamics features as the atomistic MD simulations. The force-field of the CABS model is not limited to native-like interactions, and, therefore, the results obtained in folding simulations are not assumed *a priori*. It should be noted, that in the case of the restrained simulations (described above) a part of the protein residues forming native-like interactions were weakly restrained toward their native distance (those between or within secondary structure elements), while the rest of them remained completely unrestrained (those between or within loops or between loops and secondary-structure elements).

**3.2. MD and CABS Convergence.** In addition to residue mobility analysis, we provide below further convergence



dx.doi.org/10.1021/ct300854w | J. Chem. Theory Comput. 2013, 9, 119−125

Journal of Chemical Theory and Computation Article

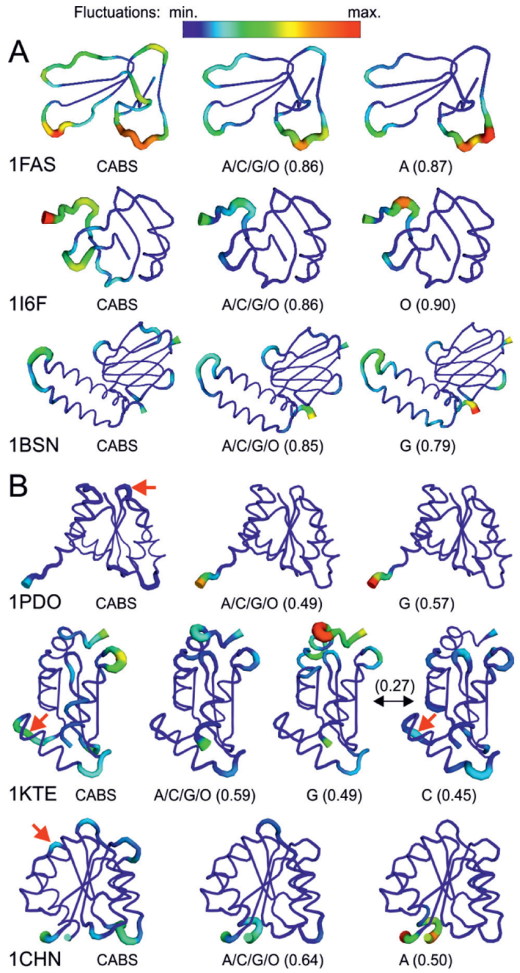



**Figure 3.** Example mobility profiles visualized on 3D structures. Profiles are shown for the three proteins with the highest (A) and the three with the lowest (B) correlation values between CABS and MD. For each protein mobility profiles are presented from the CABS model, the four MD force-fields (averaged mobility profile marked as A/C/G/O) and a single MD force-field (A, C, G, or O - always the one which yielded the highest fluctuation value for any single residue). Correlation coefficients for residue mobility between CABS and presented MD simulation are given in brackets. Colors denote fluctuation values scaled from the maximum (red) to minimum fluctuation value (blue) observed in any of the simulations. Protein chain thickness indicates the largest (thick tube) and smallest fluctuations (thin tube), for the given simulation only. Additionally, for the weakest correlation cases (B) protein fragments with the largest contribution to CABS and MD fluctuation inconsistency are marked with red arrows. For 1KTE, an additional fourth fluctuation profile is shown (from C simulations) to indicate significant inconsistency between G and C simulations ($r_s$ = 0.27) and consistency between CABS and C simulations in the marked region. The correlation coefficients and RMSD for the whole test set are given in Table S3.

analysis (for the optimal CABS setup) with different metrics for trajectory comparison. The metrics applied here are the same, or similar, as those used in the study investigating A, C, G, and O force-fields consistency.[16]

The global similarity between the structures obtained by different MD force-fields and CABS is shown in Table 2 according the similarity index Ω. The analysis shows that all simulations produce a similar picture of protein structural

**Table 2. Ω Similarity Index between MDs (A, C, G, O) and CABS Simulations[a]**

| | A | C | G | O |
|---|---|---|---|---|
| **CABS** | **0.6** | **0.6** | **0.6** | **0.6** |
| A | 1.0 | 0.7 | 0.6 | 0.7 |
| C | | 1.0 | 0.6 | 0.7 |
| G | | | 1.0 | 0.7 |

[a]Ω = 1 indicates that the simulations sample identical conformational space (in terms of pair-cross RMSD between trajectory structures), while Ω close to zero means that structural diversity is very high.

diversity, with CABS and G-simulations being slightly less similar to others than A, C, and O simulations to each other. The average effective distance ($\Omega^{-1}$) between pairs of A, C, and O simulations is around 1.4 Å, while that between pairs of CABS and G-simulations with others is around 1.7 Å . The examination of average divergences between different types of simulations ($\alpha_{AB}$ in eq 3) shows that the largest deviations are found between CABS and MD simulations (3.3−3.5 Å), while among MD force-fields the divergences are in the range of 2.2−2.9 Å (the largest for G-simulations).

The CABS trajectories appeared to be different from the different MDs and most similar to G by way of the average Lindemann's disorder index. The index provides a global measure of protein flexibility compared with that of macroscopic solids or liquids[38] (see eq 5). The average $\Delta_L$ indexes are as follows: 0.21 ± 0.03 for O; 0.22 ± 0.03 for A, C; 0.24 ± 0.03 for G; and 0.26 ± 0.03 for CABS trajectories. The slight difference in the calculated $\Delta_L$ compared to the data presented in ref 16 (average $\Delta_L$ = 0.28 ± 0.06) may result from considering only $C^\alpha$ atoms here, with more flexible portions of proteins (such as exposed side chains for which the $\Delta_L$ found[16] was 0.38 ± 0.07) being neglected.

Furthermore, we computed the overlap of deformation space (indicating similarity between the modes of two trajectories) using $\gamma$ and s overlap (see eq 6 and eq 7). The similarity indexes presented in Table 3 indicate the same level of

**Table 3. Average Deformation Space Overlaps $\gamma$ (First Number) and $s$ (After the Slash Number) between MDs (A, C, G, O) and CABS Simulations[a]**

| | A | C | G | O |
|---|---|---|---|---|
| **CABS** | **0.6/0.3** | **0.6/0.4** | **0.6/0.4** | **0.6/0.4** |
| A | 1.0/1.0 | 0.6/0.4 | 0.6/0.3 | 0.7/0.4 |
| C | | 1.0/1.0 | 0.6/0.4 | 0.7/0.4 |
| G | | | 1.0/1.0 | 0.6/0.4 |

[a]Similarity index $\gamma$ was computed for the minimum number of eigenvectors required to explain the 90% of variance. Note that when the compared trajectories span the same conformational space, the overlap value is equal 1; when the overlap value is zero, the sampled spaces are completely orthogonal ($\gamma$ and $s$ indices are described in the Materials and Methods, see eqs 6 and 7).

deformation space overlap between CABS and MD as among different MDs. The complexity of the deformability space (measured as the minimum number of eigenvectors needed to explain 90% of total variance) is higher in the case of CABS simulations than different MDs (see Figure 4). This is a similar observation to that shown in the study of coarse-grained BD and DMD models (already mentioned above), indicating that essential movements from coarse grained models might not be accurate individually, but when considered together (in the


dx.doi.org/10.1021/ct300854w | J. Chem. Theory Comput. 2013, 9, 119−125

Journal of Chemical Theory and Computation Article

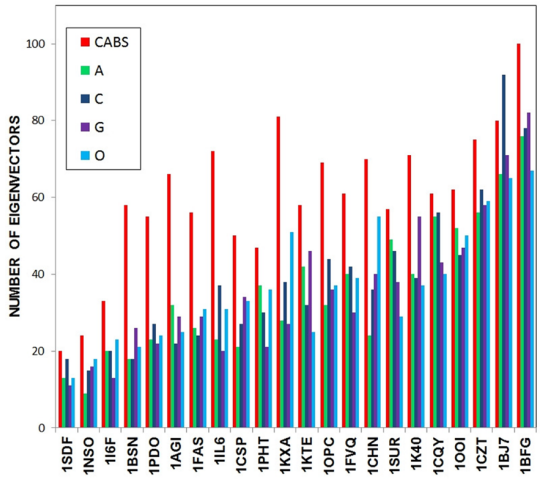



**Figure 4.** Number of essential modes required to explain 90% of variance (for each protein from the set), using CABS (shown in red) and different MDs (A, C, G, O - shown in green, blue, purple, and cyan, respectively). The proteins are listed according to the average value of essential modes for A, C, G, and O.

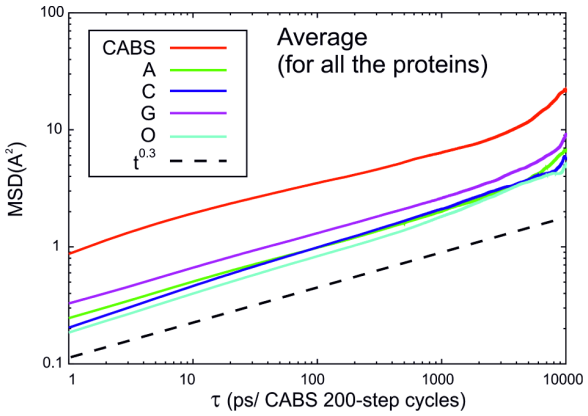



**Figure 5.** Autocorrelation function - mean square displacement (MSD) (see eq 4) of all protein residues (log scale) at different time intervals, averaged over all proteins studied. A single time unit on abscissa corresponds to 1 ps in MD simulations and 1 CABS time unit (time interval between each frame of the CABS trajectories, set to 200 MC CABS macrocycles).

essential deformation space) they provide a similar description to that obtained by MD.[45] Interestingly, the similarity index $\gamma$ between MD and CABS observed in our study (for 90% of the essential space, Table 3) is on a similar level but slightly higher (0.59) than the same index between MD and coarse grained BD and DMD models (0.51 and 0.55, respectively) observed in the Emperador et al.[45] study.

**3.3. Diffusion Properties.** Protein folding dynamics can be described as diffusion on a free energy landscape (considered as a function of one or a few chosen reaction coordinates).[48] Diffusive dynamics is characterized by mean square displacement (MSD) linearly growing with time $\langle \Delta x^2(t) \rangle = 2Dt^{\alpha}$, where $\alpha = 1$ and $D$ is the diffusion coefficient. The nonlinear relationship with time is described as "anomalous diffusion". $\alpha$ exponent values <1 and >1 correspond to subdiffusion and superdiffusion, respectively. Subdiffusion indicates that a system is trapped in local minima and the dynamics "has memory", whereas superdiffusion denotes long jumps of a system in conformational space. We studied the diffusion properties with the MSD autocorrelation function (see eq 4) to compare MD and CABS dynamics. The MSDs of all MD trajectories exhibit a power law dependence on time, with an exponent of around 0.3, just as in the CABS model (see Figure 5). This suppressed diffusion is a consequence of the relatively short time scale of the MD trajectories (the proteins are trapped in their near-native minimum) in atomic MD simulations and soft restraints imposed on protein structures (which force near-native trapping) in the case of CABS modeling.

## 4. CONCLUSIONS

A great effort has been expended in recent decades to the quest for efficient and accurate algorithms for protein dynamics simulation. Numerous methods have been exercised utilizing various sampling, representation, and force-field models. Atomistic MD, employing Newton's laws of motion and empirical energy functions, emerged as gold standard of protein dynamics simulations. Apart from the improvement of many problems of classical MD techniques, current research seeks for novel computational approaches capable of moving protein simulations to higher coverage of conformational space and better accuracy. We attempted to examine and maximize the consistency of short-time protein dynamics by atomistic MD and the CABS model, two qualitatively different approaches with respect to sampling, representation, and force-field. Considering the conceptual difference between the methods, they both offer a surprisingly similar picture of protein structure flexibility (the average Spearman's correlation coefficient for the fluctuations along protein chains from the protein set is 0.7).

This work offers promising prospects for the following: (1) fast prediction of MD results by the CABS model (for at least short time scale dynamics) and (2) bridging the CABS and atomistic MD into a single multiscale protocol for the simulation of protein dynamics in atomic resolution (in which MD could be bootstrapped from representative models from the CABS dynamics). Development of such multiscale procedures offers simulation approach of similar quality to atomic MD but many times faster. We roughly estimate CABS dynamics to be $6 \times 10^3$ cheaper in terms of computational cost than the classical MD (based on Rueda et al.[16] estimations giving on average 3650 CPU hours for single protein simulation from a test set, compared to 0.6 CPU hour by the CABS model).

## ■ ASSOCIATED CONTENT

### Supporting Information

Tables: S1 (protein data set), S2 (the five top ranked parameters setups and respective $r_s$ values for the training, the test, and the whole protein set), S3 (Spearman's correlation coefficients and mean RMSD (after the slash) between MDs (A, C, G, O) and CABS mobility profiles), and Figures: S1 (restraints map for 1I6F), S2 (mobility profiles by the CABS model and MD force-fields, for the protein test set). This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: sekmi@chem.uw.edu.pl.

### Notes

The authors declare no competing financial interest.

Journal of Chemical Theory and Computation Article

## ■ ACKNOWLEDGMENTS

The authors thank Dr. Tim Meyer for critical reading of the manuscript. We would like to acknowledge support from a Project operated within the Foundation for Polish Science MPD Programme and TEAM project (TEAM/2011-7/6) cofinanced by the EU European Regional Development Fund operated within the Innovative Economy Operational Program, and from Polish National Science Center (Grant No. NN301071140), and from Polish Ministry of Science and Higher Education Grant No. IP2011 024371.

## ■ REFERENCES

(1) Russel, D.; Lasker, K.; Phillips, J.; Schneidman-Duhovny, D.; Velazquez-Muriel, J. A.; Sali, A. *Curr. Opin. Cell Biol.* **2009**, *21*, 97−108.
(2) Kmiecik, S.; Jamroz, M.; Kolinski, A. In *Multiscale Approaches to Protein Modeling*; Kolinski, A., Ed.; Springer: New York, 2011; Chapter 12, pp 281−294.
(3) Lin, M. M.; Mohammed, O. F.; Jas, G. S.; Zewail, A. H. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 16622−16627.
(4) Zhang, J.; Baker, M. L.; Schroder, G. F.; Douglas, N. R.; Reissmann, S.; Jakana, J.; Dougherty, M.; Fu, C. J.; Levitt, M.; Ludtke, S. J.; Frydman, J.; Chiu, W. *Nature* **2010**, *463*, 379−383.
(5) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94*, L75−L77.
(6) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341−346.
(7) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517−520.
(8) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47−L49.
(9) Freddolino, P. L.; Schulten, K. *Biophys. J.* **2009**, *97*, 2338−2347.
(10) Ensign, D. L.; Kasson, P. M.; Pande, V. S. *J. Mol. Biol.* **2007**, *374*, 806−816.
(11) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712−725.
(12) Mittal, J.; Best, R. B. *Biophys. J.* **2010**, *99*, L26−L28.
(13) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. *J. Mol. Biol.* **2011**, *405*, 43−48.
(14) Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011−19016.
(15) Ensign, D. L.; Pande, V. S. *Biophys. J.* **2009**, *96*, L53−L55.
(16) Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Perez, A.; Camps, J.; Hospital, A.; Gelpi, J. L.; Orozco, M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 796−801.
(17) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.
(18) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(19) Mackerell, A. D.; Wiorkiewiczkuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946−11975.
(20) MacKerell, A. D.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(21) Ott, K. H.; Meyer, B. *J. Comput. Chem.* **1996**, *17*, 1068−1084.
(22) Hermans, J.; Berendsen, H. J. C.; Vangunsteren, W. F.; Postma, J. P. M. *Biopolymers* **1984**, *23*, 1513−1518.
(23) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(24) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.
(25) Kmiecik, S.; Kolinski, A. *Biophys. J.* **2008**, *94*, 726−736.
(26) Kmiecik, S.; Gront, D.; Kolinski, A. *BMC Struct. Biol.* **2007**, *7*, 1−11.
(27) Kmiecik, S.; Gront, D.; Kouza, M.; Kolinski, A. *J. Phys. Chem. B* **2012**, *116*, 7026−7032.
(28) Kmiecik, S.; Kolinski, A. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 12330−12335.
(29) Kmiecik, S.; Kolinski, A. *J. Am. Chem. Soc.* **2011**, *133*, 10283−10289.
(30) Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Pérez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; Gelpí, J. L.; Orozco, M. *Structure* **2010**, *18*, 1399−1409.
(31) Kolinski, A.; Skolnick, J. *Polymer* **2004**, *45*, 511−524.
(32) Liwo, A.; He, Y.; Scheraga, H. A. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16890−16901.
(33) Scheraga, H. A.; Khalili, M.; Liwo, A. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57−83.
(34) Kolinski, A. *Acta Biochim. Pol.* **2004**, *51*, 349−371.
(35) Kolinski, A.; Bujnicki, J. M. *Proteins* **2005**, *61*, 84−90.
(36) Jamroz, M.; Kolinski, A. *BMC Struct. Biol.* **2010**, *10*, 1−9.
(37) Gront, D.; Kolinski, A. *Bioinformatics* **2008**, *24*, 584−585.
(38) Zhou, Y.; Vitkup, D.; Karplus, M. *J. Mol. Biol.* **1999**, *285*, 1371−1375.
(39) Meyer, T.; Ferrer-Costa, C.; Perez, A.; Rueda, M.; Bidon-Chanal, A.; Luque, F. J.; Laughton, C. A.; Orozco, M. *J. Chem. Theory Comput.* **2006**, *2*, 251−258.
(40) Hess, B. *Phys. Rev. E* **2000**, *62*, 8438−8448.
(41) Hess, B. *Phys. Rev. E* **2002**, *65*, 031910.
(42) Jamroz, M.; Kolinski, A.; Kihara, D. *Proteins* **2012**, *80*, 1425−1435.
(43) Yang, L.; Song, G.; Jernigan, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 12347−12352.
(44) Emperador, A.; Meyer, T.; Orozco, M. *J. Chem. Theory Comput.* **2008**, *4*, 2001−2010.
(45) Emperador, A.; Carrillo, O.; Rueda, M.; Orozco, M. *Biophys. J.* **2008**, *95*, 2127−2138.
(46) McCammon, J. A.; Harvey, S. C. *Dynamics of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, UK, 1987.
(47) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459−466.
(48) Krivov, S. V. *PLoS Comput. Biol.* **2010**, *6*, e1000921.

Supporting Information
*"A Consistent View of Protein Fluctuations from All-atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-based Force-field"*

**Table S1.** Protein data set.

| PDB ID code | No. of domains and CATH classification (Class, Architecture) | No. of residues | No. of disulfide bridges |
|---|---|---|---|
| 1AGI | 1, Alpha/Beta, roll | 125 | 3 |
| 1BFG | 1, mainly Beta, trefoil | 126 | 0 |
| 1BJ7 | 1, mainly Beta, beta barrel | 150 | 2 |
| 1BSN | 2, mainly Beta, sandwich and mainly Alpha, up-down bundle | 138 | 0 |
| 1CHN | 1, Alpha/Beta, 3-layer(aba) sandwich | 126 | 0 |
| 1CQY | 1, mainly Beta (sandwich) | 99 | 0 |
| 1CSP | 1, mainly Beta (beta barrel) | 67 | 0 |
| 1CZT | 1, mainly Beta, sandwich | 160 | 1 |
| 1FAS | 1, mainly Beta, ribbon | 61 | 4 |
| 1FVQ | 1, Alpha/Beta, 2-layer Sandwich | 72 | 0 |
| 1I6F | 1, Alpha/Beta, 2-layer Sandwich | 60 | 4 |
| 1IL6 | 1, mainly Alpha, up-down Bundle | 166 | 4 |
| 1K40 | 1, mainly Alpha, up-down Bundle | 126 | 0 |
| 1KTE | 1, Alpha/Beta, 3-layer(aba) sandwich | 105 | 1 |
| 1KXA | 2, mainly Beta, beta barrel (both domains) | 158 | 0 |
| 1NSO | 1, mainly Beta, beta barrel | 107 | 0 |
| 1OOI | 1, mainly Alpha, orthogonal bundle | 124 | 3 |
| 1OPC | 1, mainly Alpha, orthogonal bundle | 99 | 0 |
| 1PDO | 1, Alpha/Beta, 3-layer(aba) sandwich | 129 | 0 |
| 1PHT | 1, mainly Beta, roll | 83 | 0 |
| 1SDF | 1, mainly Beta, beta Barrel | 67 | 2 |
| 1SUR | 1, Alpha/Beta, 3-layer(aba) sandwich | 215 | 0 |

For each protein, the following data are presented: Protein Data Bank code, number of domains, 2 levels of CATH classification (class and architecture) number of residues and number of disulfide bridges.

**Optimization procedure of CABS parameters for simulations with distance restraints and predictive power test**

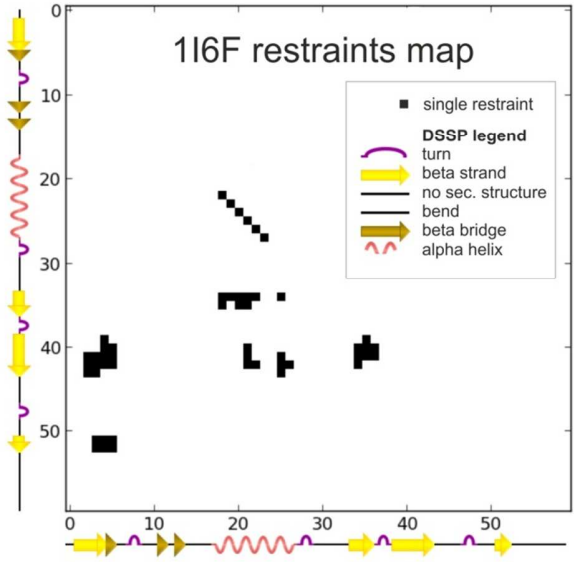
The following CABS parameters have been tested in order to obtain the highest possible convergence with the MD mobility profiles (tested values are given in the brackets):

- l1 - simulation length parameter 1 (200, 1000, 2000, 5000, 10000) – defines number of CABS MC macrocycles
- l2 - simulation length parameter 2 (1, 10, 50, 100) – determines intervals length between recorded snapshots
- t - reduced temperature (0.8, 1.0, 1.2, 1.4, 1.6, 1.8)
- rst - restraints strength (1, 2, 3, 5) – a force constant that determines the magnitude of the energy penalty for the deviation of a residue position from $x \pm \Delta$ (see $\Delta$ description below). The potential for the constraint is flat bottom with linear sides. The energy penalty increases from zero as the distance deviates from $x \pm \Delta$ and grows linearly with the distance (the force constant is the slope).

- Δ - defines restraints range as [x-Δ,x+Δ] in Angstroms (0.5, 1, 2, 3), where x is the native distance between restrained residues
- rt - restraint type - 4 versions with respect to range type (global, local) and density (high, low) were tested: global-high, global-low, local-high, local-low.

The CABS parameters were tuned iteratively by running multiple parameter set-ups (on the whole protein test) and expert-guided modifications between the iterations (the expert interference was based on exclusion of the parameters options which clearly worsened the convergence of the fitting procedure, as well as on expanding the parameters options to be tested). The optimal parameters setup, which yielded overall the highest mean correlations (on the level of 0.74), were obtained with l1=10000, l2=10, t=1.2, rst=2, Δ=1, and weak native-like restraints of low density, applied only locally and between residues belonging to the secondary structure elements (alpha or beta) (for the example restraint map see Figure S1). On average, 125 restraints were used per one protein (number of restraints for a single protein varied from 45 to 220 depending on the type of secondary structure motifs and protein length).

**Figure S1.** Restraints map for 1I6F (an example of alpha/beta protein from the test set). Every black square denote a single distance restraint (between respective c-alpha atoms) applied in the optimal simulation set-up. Secondary structure elements, assigned according to DSSP algorithm, are marked on the map borders. 1I6F is a 60-residue protein and 50 restraints were applied.




Furthermore, in order to test the predictive power of the method, we randomly split the protein set into a training and a test set. The method was parameterized based on the training set only and subsequently tested on the test set. The results for the top five parameters set-ups, ranked according to the average Spearman's correlation coefficient value ($r_s$) for the training set (and respective $r_s$ values for the test set and for the entire set), are listed in the Table S2.

**Table S2.** The five top ranked parameters set-up's and respective $r_s$ values for the training, the test and the whole protein set.

| $r_s$ | | | Parameters | | | | | |
|---|---|---|---|---|---|---|---|---|
| Training set | Test set | Overall | l1 | l2 | t | rst | Δ | rt |
| 0.78 | 0.70 | 0.74 | 10000 | 10 | 1.2 | 2 | 1 | loc.-low |
| 0.76 | 0.71 | 0.74 | 5000 | 10 | 1.4 | 2 | 1 | loc.-low |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.76 | 0.71 | 0.73 | 1000 | 10 | 1.4 | 2 | 1 | loc.-low |
| 0.76 | 0.70 | 0.73 | 10000 | 10 | 1.2 | 3 | 1 | loc.-low |
| 0.75 | 0.72 | 0.73 | 10000 | 10 | 1.4 | 2 | 1 | loc.-low |

The randomly chosen training set included the following proteins: 1K40, 1KTE, 1CZT, 1FAS, 1OPC, 1SUR, 1IL6, 1KXA, 1CHN, 1SDF, 1I6F, while the test set: 1CSP, 1BJ7, 1BFG, 1BSN, 1PDO, 1NSO, 1PHT, 1FVQ, 1AGI, 1CQY, 1OOI.

The method trained on the training set showed the highest mean $r_s$ for the same parameters setup as trained on the whole protein set and predictive power on the level of 0.7 ($r_s$ for the test set). The parameterization results for the top five ranked parameters set-ups (Table S2) shows that similarly good results can be obtained from 2 or 10 times shorter simulation runs (l1=5000, l2=1000), in the narrow temperature range (t=1.2 or t=1.4), while the other parameters found should be preferably fixed.

**Table S3.** Spearman's correlation coefficients and mean RMSD (after the slash) between MD's (A, C, G, O) and CABS mobility profiles.

| | O/G | A/C | O/C | O/A | A/G | G/C | **A/CABS** | **C/CABS** | **G/CABS** | **O/CABS** | **all MD/ CABS** | **all MD/ CABS-U** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1AGI** | 0.68 / 4.70 | 0.77 / 2.54 | 0.82 / 2.79 | 0.75 / 1.63 | 0.65 / 3.48 | 0.70 / 4.80 | 0.78 / 3.63 | 0.81 / 4.80 | 0.72 / 1.43 | 0.71 / 4.70 | **0.76 / 3.64** | 0.74 / 7.07 |
| **1BFG** | 0.79 / 0.78 | 0.87 / 0.38 | 0.87 / 0.39 | 0.84 / 0.24 | 0.80 / 0.71 | 0.78 / 0.73 | 0.74 / 1.91 | 0.77 / 1.91 | 0.65 / 1.58 | 0.73 / 2.00 | **0.72 / 1.85** | 0.64 / 6.10 |
| **1BJ7** | 0.81 / 0.82 | 0.84 / 0.52 | 0.89 / 0.76 | 0.80 / 0.77 | 0.71 / 1.19 | 0.73 / 1.33 | 0.70 / 2.96 | 0.66 / 3.20 | 0.63 / 2.43 | 0.68 / 2.93 | **0.67 / 2.88** | 0.71 / 4.88 |
| **1BSN** | 0.66 / 3.83 | 0.80 / 3.28 | 0.81 / 2.94 | 0.80 / 2.43 | 0.77 / 2.82 | 0.71 / 4.15 | 0.85 / 2.58 | 0.88 / 2.06 | 0.79 / 3.56 | 0.87 / 2.52 | **0.85 / 2.68** | 0.80 / 5.59 |
| **1CHN** | 0.75 / 1.46 | 0.82 / 3.42 | 0.85 / 1.15 | 0.77 / 4.03 | 0.74 / 3.29 | 0.82 / 1.28 | 0.50 / 3.65 | 0.64 / 1.93 | 0.74 / 2.13 | 0.68 / 1.99 | **0.64 / 2.43** | 0.44 / 5.59 |
| **1CQY** | 0.86 / 1.12 | 0.75 / 0.50 | 0.85 / 0.88 | 0.86 / 0.93 | 0.90 / 1.39 | 0.79 / 1.32 | 0.70 / 2.74 | 0.70 / 2.70 | 0.75 / 1.92 | 0.78 / 2.63 | **0.73 / 2.50** | 0.52 / 4.40 |
| **1CSP** | 0.89 / 1.70 | 0.91 / 0.51 | 0.86 / 0.91 | 0.92 / 0.75 | 0.82 / 1.37 | 0.84 / 1.11 | 0.84 / 1.42 | 0.78 / 1.30 | 0.63 / 1.50 | 0.73 / 1.49 | **0.75 / 1.43** | 0.70 / 4.49 |
| **1CZT** | 0.83 / 1.16 | 0.94 / 0.67 | 0.88 / 0.66 | 0.92 / 0.48 | 0.89 / 1.28 | 0.89 / 1.31 | 0.88 / 3.12 | 0.83 / 3.19 | 0.83 / 2.99 | 0.86 / 3.09 | **0.85 / 3.10** | 0.78 / 4.71 |
| **1FAS** | 0.88 / 1.00 | 0.79 / 0.78 | 0.85 / 0.68 | 0.89 / 0.33 | 0.90 / 1.09 | 0.80 / 0.86 | 0.87 / 1.53 | 0.82 / 1.24 | 0.91 / 0.88 | 0.84 / 1.45 | **0.86 / 1.28** | 0.81 / 4.51 |
| **1FVQ** | 0.64 / 0.84 | 0.75 / 0.40 | 0.77 / 0.64 | 0.74 / 0.61 | 0.83 / 0.72 | 0.69 / 0.81 | 0.67 / 0.77 | 0.72 / 0.74 | 0.66 / 0.78 | 0.81 / 0.65 | **0.72 / 0.74** | 0.35 / 9.06 |
| **1I6F** | 0.81 / 3.20 | 0.86 / 1.23 | 0.85 / 2.61 | 0.89 / 2.75 | 0.86 / 1.30 | 0.73 / 1.93 | 0.87 / 4.67 | 0.88 / 4.16 | 0.79 / 5.00 | 0.90 / 3.82 | **0.86 / 4.41** | 0.72 / 7.11 |
| **1IL6** | 0.80 / 4.17 | 0.84 / 4.72 | 0.82 / 2.47 | 0.89 / 4.19 | 0.77 / 5.05 | 0.85 / 3.22 | 0.80 / 4.90 | 0.87 / 2.57 | 0.85 / 2.61 | 0.82 / 3.46 | **0.84 / 3.39** | 0.75 / 5.92 |
| **1K40** | 0.79 / 1.95 | 0.81 / 1.17 | 0.82 / 1.32 | 0.89 / 0.70 | 0.68 / 1.73 | 0.74 / 0.81 | 0.79 / 0.79 | 0.75 / 1.52 | 0.73 / 2.12 | 0.88 / 0.83 | **0.79 / 1.32** | 0.64 / 4.24 |
| **1KTE** | 0.62 / 2.87 | 0.49 / 0.96 | 0.62 / 1.22 | 0.86 / 0.84 | 0.54 / 2.92 | 0.26 / 2.85 | 0.76 / 1.93 | 0.45 / 1.64 | 0.49 / 2.70 | 0.66 / 2.14 | **0.59 / 2.10** | 0.44 / 5.16 |
| **1KXA** | 0.87 / 4.39 | 0.78 / 3.22 | 0.86 / 1.97 | 0.88 / 2.40 | 0.83 / 5.23 | 0.82 / 3.38 | 0.80 / 2.76 | 0.77 / 2.89 | 0.80 / 4.36 | 0.82 / 2.60 | **0.80 / 3.15** | 0.73 / 6.27 |
| **1NSO** | 0.56 / 6.63 | 0.55 / 5.13 | 0.61 / 4.80 | 0.68 / 4.95 | 0.63 / 7.22 | 0.54 / 5.51 | 0.53 / 6.36 | 0.75 / 4.71 | 0.60 / 4.90 | 0.66 / 6.69 | **0.67 / 5.67** | 0.69 / 7.61 |
| **1OOI** | 0.70 / 0.74 | 0.78 / 0.69 | 0.85 / 0.79 | 0.81 / 0.71 | 0.67 / 0.81 | 0.69 / 0.70 | 0.71 / 2.11 | 0.61 / 1.87 | 0.67 / 1.85 | 0.66 / 2.15 | **0.66 / 2.00** | 0.20 / 4.15 |
| **1OPC** | 0.59 / 1.75 | 0.80 / 1.16 | 0.69 / 1.19 | 0.60 / 1.46 | 0.87 / 1.04 | 0.86 / 1.15 | 0.63 / 1.90 | 0.72 / 1.88 | 0.64 / 1.87 | 0.71 / 1.85 | **0.68 / 1.88** | 0.57 / 4.38 |
| **1PDO** | 0.72 / 3.21 | 0.84 / 2.19 | 0.91 / 1.23 | 0.83 / 2.38 | 0.80 / 4.58 | 0.83 / 4.17 | 0.53 / 3.88 | 0.51 / 4.30 | 0.57 / 7.65 | 0.34 / 5.14 | **0.49 / 5.24** | 0.64 / 6.38 |
| **1PHT** | 0.70 / 1.72 | 0.90 / 1.35 | 0.76 / 1.44 | 0.80 / 1.59 | 0.79 / 0.82 | 0.75 / 1.62 | 0.70 / 4.48 | 0.73 / 3.55 | 0.68 / 4.40 | 0.79 / 3.42 | **0.73 / 3.96** | 0.73 / 6.61 |
| **1SDF** | 0.87 / 2.30 | 0.83 / 2.11 | 0.87 / 2.96 | 0.92 / 2.80 | 0.87 / 3.28 | 0.83 / 2.80 | 0.81 / 7.09 | 0.78 / 6.81 | 0.83 / 4.62 | 0.79 / 5.16 | **0.80 / 5.92** | 0.51 / 8.32 |
| **1SUR** | 0.76 / 3.05 | 0.81 / 1.52 | 0.72 / 1.14 | 0.81 / 1.77 | 0.86 / 3.56 | 0.82 / 3.22 | 0.84 / 3.47 | 0.78 / 3.48 | 0.81 / 2.79 | 0.73 / 3.47 | **0.79 / 3.30** | 0.73 / 5.22 |
| **AVERAGE** | **0.75 / 2.43** | **0.80 / 1.75** | **0.81 / 1.59** | **0.82 / 1.76** | **0.78 / 2.49** | **0.75 / 2.23** | **0.74 / 3.12** | **0.74 / 2.84** | **0.72 / 2.91** | **0.75 / 2.92** | **0.74 / 2.95** | **0.63 / 5.81** |

Each table cell contains appropriate Spearman's correlation value for residue mobility and the mean RMSD (Å) value (RMSD between mobility profiles) between compared trajectories. Next to last column (all MD/CABS) shows the averaged values between all MD force fields and the CABS, the last column (all MD/CABS-U) shows the averaged values between all MD and the CABS unrestrained simulations.

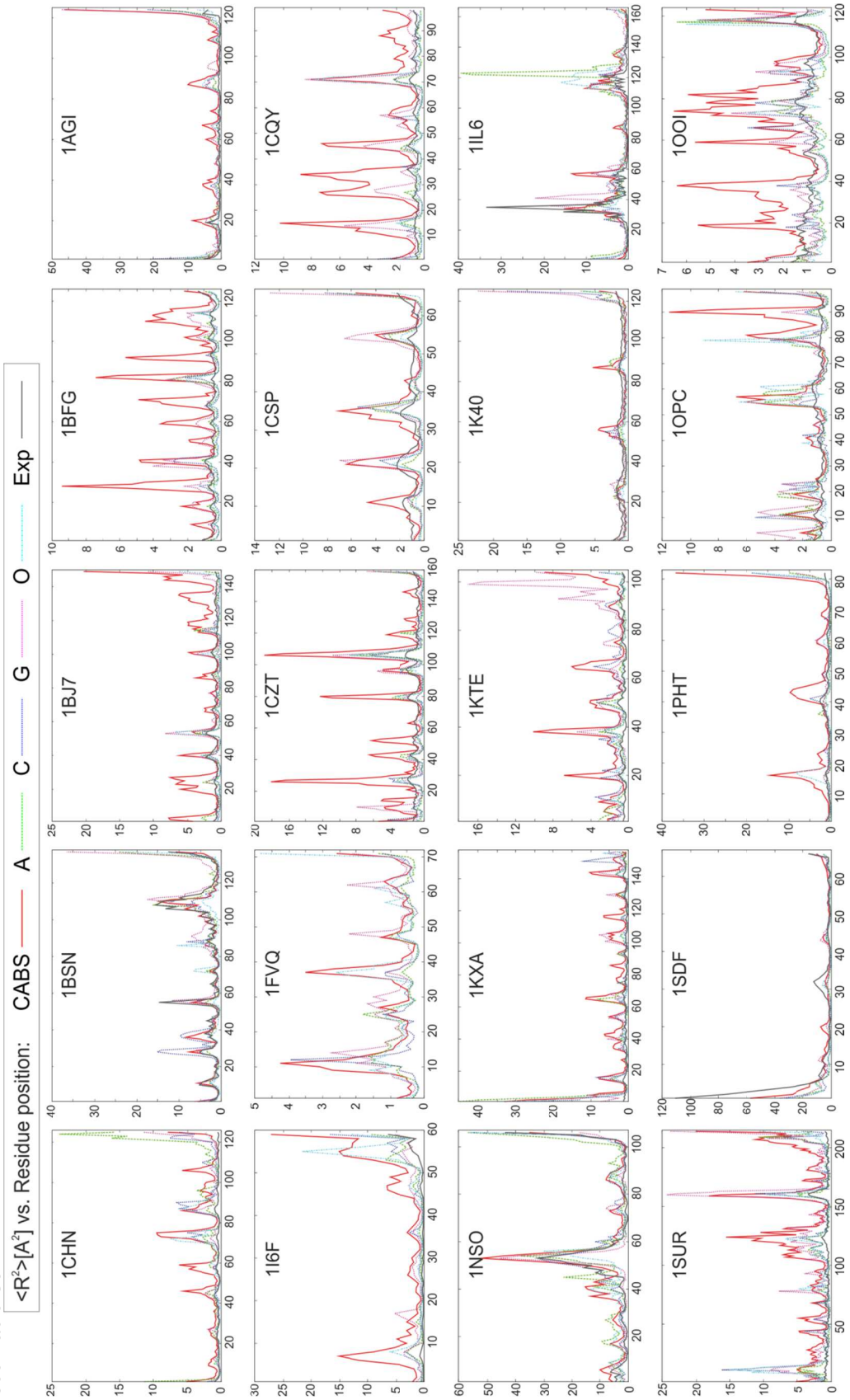



**Figure S2.** Mobility profiles by the CABS model, MD force-fields and experimental data (obtained from crystallographic B-factors or NMR multiple structures) for the protein test set (1FAS and 1PDO profiles are shown in Figure 2). For Spearman's correlation coefficient values of each protein, see Table S3.

# Dodatki

# F MASZYNY WEKTORÓW NOŚNYCH, SVM

Maszyny wektorów nośnych (ang.: *Support Vector Machine*) (Cortes i Vapnik, 1995) są stosunkowo nową metodą uczenia maszynowego, mającą zastosowanie w wielu dziedzinach nauki — od bioinformatyki po ekonomię.

W pierwotnej wersji metoda służyła do klasyfikacji danych, tj. znalezienia hiperpłaszczyzny dzielącej zbiór danych $(\mathbf{x_i}, y_i)$ (gdzie $\mathbf{x_i} \in \mathbb{R}^p$) tak, by otrzymać dwie grupy, dla których $y_i \in \pm 1$. W istocie przestrzeń dzieli się dwiema hiperpłaszczyznami, oddzielonymi od siebie o margines błędu $2\epsilon$.

W podejściu regresji wektorów nośnych (SVR) stosuje się dość zaawansowany aparat matematyczny pozwalający na wyznaczenie takiej hiperpłaszczyzny, dla której większość punktów $\mathbf{x}$ leżeć będzie w granicy marginesu błędu, $\epsilon$.

# G MIARY (NIE)PODOBIEŃSTWA UŻYWANE W PRACY

W rozdziale tym przedstawię miary podobieństwa (lub niepodobieństwa), którymi posługiwałem się w trakcie badań, lecz nie zostały one szczegółowo opisane w publikacjach stanowiących podstawę rozprawy.

## G.1 WSPÓŁCZYNNIK KORELACJI PEARSONA

Współczynnik ten, wyrażony równaniem:

$$r_{xy} \equiv \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^N (x_i - \overline{x})^2}\sqrt{\sum_i^N (y_i - \overline{y})^2}} \tag{G.1}$$

gdzie:

$r_{xy}$ współczynnik korelacji Pearsona pomiędzy zbiorami N punktów $(x_i, y_i)$. $r_{xy} \in [-1, 1]$,

$\overline{x}$ wartość średnia ze zbioru $x$,

określa liniową zależność zmiennych $y$ od $x$. Gdy $r_{xy} = 1$, dane są dodatnio zupełnie skorelowane, tj. można wyznaczyć prostą $y = ax + b$ przechodzącą przez wszystkie punkty $(x, y)$. Gdy $r_{xy} = -1$, dane są ujemnie zupełnie skorelowane, tj. można wyznaczyć prostą $y = -ax + b$ przechodzącą przez wszystkie punkty.

## G.2 WSPÓŁCZYNNIK KORELACJI SPEARMANA

Współczynnik korelacji Spearmana (korelacja rangowa) pozwala na sprawdzenie, czy zbiór punktów $(x_i, y_i)$ związany jest funkcją monotoniczną. Wprowadza się tu ranking zmiennych $x$ oraz $y$ (ułożenie zmiennych w kolejności

rosnącej i przypisanie im indeksów 1, 2, 3 ... odpowiadających pozycji w szeregu), a następnie stosuje wzór:

$$r_s \equiv 1 - 6 \sum_i^N \frac{d_i^2}{N(N^2-1)} \tag{G.2}$$

gdzie:

$r_s$ współczynnik korelacji Spearmana dla zbioru N punktów $(x_i, y_i)$, $r_s \in [-1, 1]$,

$d_i$ różnica indeksów z rankingu dla pary $(x_i, y_i)$.

Współczynnik $r_s$ jest — w porównaniu ze współczynnikiem korelacji Pearsona — mniej wrażliwy na obserwacje odstające.

## G.3 PIERWIASTEK ŚREDNIEGO KWADRATOWEGO ODCHYLENIA POŁOŻEŃ ATOMÓW, RMSD

Jest to prawdopodobnie najczęściej wykorzystywana miara porównująca podobieństwo struktur białek, wyrażona równaniem (Kabsch, 1976):

$$\mathrm{RMSD} \equiv \sqrt{\frac{1}{N} \sum_i^N \|x_i - y_i\|^2} \tag{G.3}$$

po optymalnym nałożeniu. Gdy wymagana jest jedynie wartość RMSD, nie zaś macierz obrotu, należy znaleźć m.in. wartości własne kwadratu macierzy kowariancji. W tym celu — mając dwie struktury, wyrażone wektorami $\mathbf{X}$, $\mathbf{Y}$ o wymiarach $3 \times N$ — należy (Damm i Carlson, 2006):

1. Przenieść struktury do wspólnego środka masy (bądź początku układu współrzędnych):

$$\mathbf{X}' = \mathbf{X} - \frac{\sum_i^N m_i x_i}{\sum_i^N m_i} \tag{G.4}$$

gdzie:

$\mathbf{X}'$ współrzędne struktury $\mathbf{X}$ o środku masy w początku układu współrzędnych,

$m_i$ masa atomu o współrzędnej $x_i$ (dla atomów tego samego typu można przyjąć $m = 1$).

2. Obliczyć $3 \times 3$ macierz kowariancji $\mathbf{R}$ i jej wyznacznik:

$$\mathbf{R} = \mathbf{Y}^T\mathbf{X} \tag{G.5}$$

o elementach:

$$r_{ij} = \sum_k^N y_{k,i}x_{k,j}, \tag{G.6}$$

przy założeniu, że struktury zostały przeniesione do początku układu współrzędnych, tj. $\overline{x} = 0$.

3. Obliczyć kwadrat macierzy kowariancji:

$$\mathbf{R}^2 = \mathbf{R}^T\mathbf{R}. \tag{G.7}$$

4. Wyznaczyć wartości własne $\lambda$ macierzy $\mathbf{R}^2$, np. rozwiązując układ równań trzeciego stopnia wielomianu $\det(\mathbf{R}^2 - \lambda\mathbb{I}) = 0$.

5. Obliczyć promień żyracji obu struktur (przy założeniu, że $\overline{x} = 0$ oraz $m = 1$):

$$R_{g,X}^2 = \frac{1}{N}\sum_i^N x_i^2. \tag{G.8}$$

6. Wartość RMSD otrzymuje się z równania (Brüschweiler, 2003):

$$\mathrm{RMSD} = \sqrt{R_{g,X}^2 + R_{g,Y}^2 - 2\left(\sqrt{\lambda_1} + \sqrt{\lambda_2} + s\sqrt{\lambda_3}\right)} \tag{G.9}$$

gdzie:

$\lambda_{1\ldots3}$ posortowane od największej do najmniejszej wartości własne macierzy $\mathbf{R}^2$,

$s$ znak wyznacznika macierzy kowariancji.

## G.4 GLOBAL DISTANCE TEST – TOTAL SCORE, GDT_TS

Miara RMSD, choć często stosowana, nie jest pozbawiona wad. Wspomniałem o tym już we wstępie (strona 3), sugerując, że porównywanie struktur przy założeniu, że są one sztywne (nieruchome) może prowadzić do błędów w interpretacji czy walidacji metod przewidywania struktur białek. Wychodząc naprzeciw temu problemowi, Zemla i in. (1999) zaproponowali inną miarę, GDT_TS:

$$\mathrm{GDT\_TS}_{\mathrm{opt}} \equiv \frac{1}{4}\left(\max \mathrm{C}_{1\text{Å}} + \max \mathrm{C}_{2\text{Å}} + \max \mathrm{C}_{4\text{Å}} + \max \mathrm{C}_{8\text{Å}}\right) \tag{G.10}$$

gdzie:

- $\mathrm{GDT\_TS}_{\mathrm{opt}}$ optymalna wartość miary. $\mathrm{GDT\_TS}_{\mathrm{opt}} \in [0, 1]$. Ze względu na złożoność problemu, $\mathrm{GDT\_TS} \leqslant \mathrm{GDT\_TS}_{\mathrm{opt}}$ (Li i in., 2011),
- $\max \mathrm{C}_{1\text{Å}}$ oznacza liczbę atomów (w praktyce atomów C$\alpha$) leżących w odległości nie dalszej niż 1 Å po zastosowaniu takiego nałożenia struktur, by uzyskać wartość maksymalną.

Choć w praktyce nie jest możliwe znalezienie optymalnej wartości GDT_TS (problem należy do klasy NP-trudnych), miara ta jest szeroko stosowana w trakcie walidacji wyników podczas konkursów CASP.

By obliczyć wartość GDT_TS stosuje się metody heurystyczne[1] poszukiwania takich macierzy obrotu, by zmaksymalizować wynik. Procedura postępowania może wyglądać następująco:

1. By znaleźć optymalną macierz obrotu fragmentu struktury $\mathbf{X}$ na fragment struktury $\mathbf{Y}$ należy wyznaczyć wektory własne $\mathbf{v}$ macierzy $\mathbf{R}^2_{\mathrm{frag}}$ (zdefiniowanej w Równaniu G.7), przez rozwiązanie układu równań $(\lambda\mathbb{I} - \mathbf{R}^2_{\mathrm{frag}})\mathbf{v} = 0$ przy znanych wartościach własnych (punkt 4 w Rozdziale G.3).

1 W związku z czym wartość GDT_TS może się różnić w zależności od zastosowanego algorytmu.

2. Obliczyć:

$$\mathbf{U} = (\mathbf{R}_{frag} \times \mathbf{v})^T \mathbf{v}. \tag{G.11}$$

3. Zastosować ją dla wszystkich atomów struktury $\mathbf{X}$:

$$\mathbf{X}^T_{obr.} = \mathbf{U}\mathbf{X}^T. \tag{G.12}$$

4. Policzyć odległości między parami atomów obu struktur $(x_i, y_i)$, wybierając podzbiór atomów do następnej iteracji (np. atomy leżące nie dalej niż 6 Å po nałożeniu z poprzedniego kroku (Hubbard, 1999)).

5. Zaktualizować zmienne $C := \max(C, C_{aktualne})$.

6. Zakończyć, gdy brak jest nowych podzbiorów i policzyć wartość z wykorzystaniem Równania G.10.